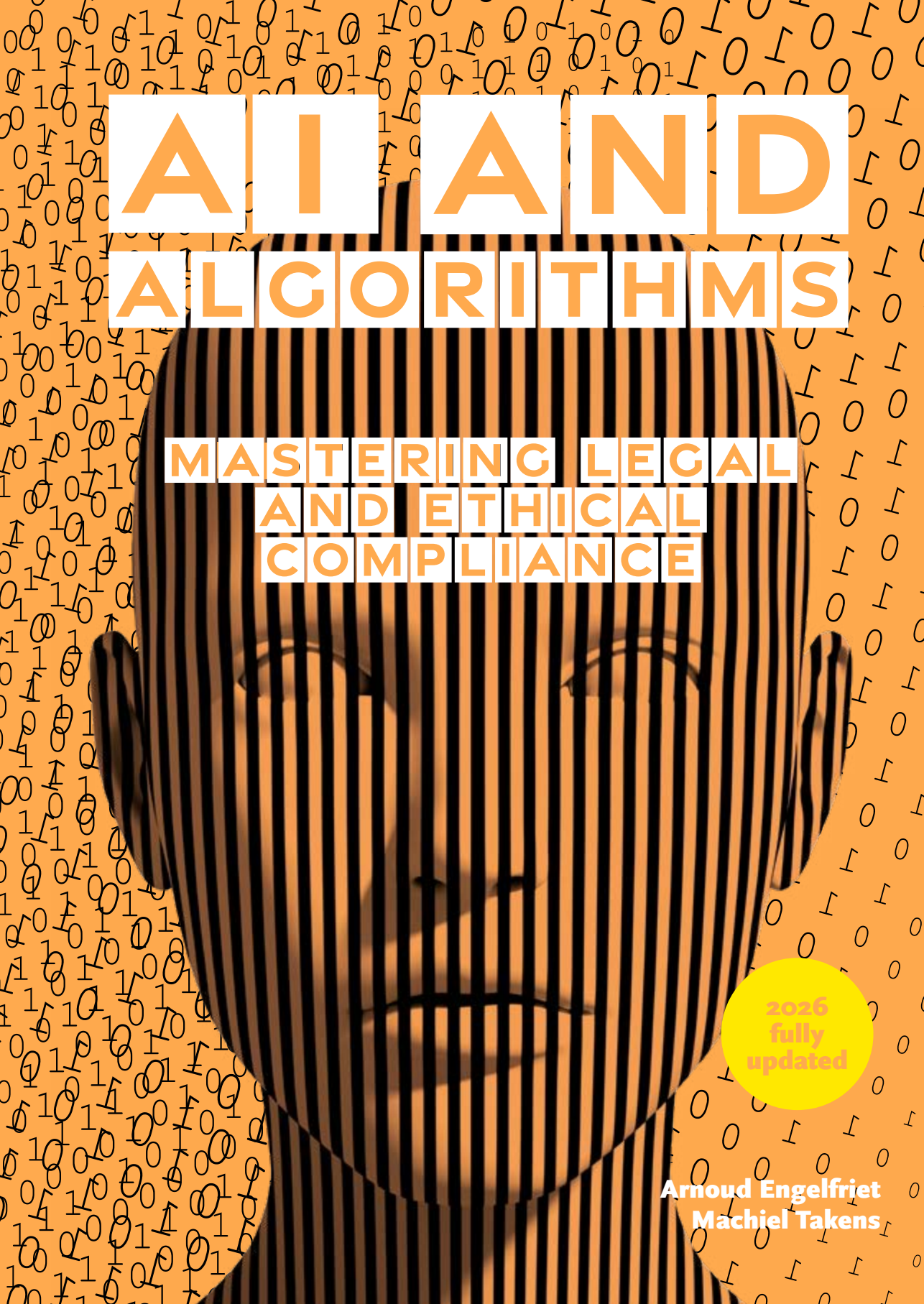
AI AND ALGORITHMS

MASTERING LEGAL AND ETHICAL COMPLIANCE



2026 fully updated

Arnoud Engelfriet
Machiel Takens

AI AND ALGORITHMS

MASTERING LEGAL AND ETHICAL COMPLIANCE

Arnoud Engelfriet
Machiel Takens

ICTRECHT

ISBN 978-90-835678-2-2

Publisher: Ius Mentis te Amsterdam
Cover, design, typesetting: Jellmedia.nl
Printing: New energy printing

TABLE OF CONTENTS

CHAPTER 3 Operationalizing AI Compliance

CHAPTER 4 Reinforcing Human Agency and Oversight

CHAPTER 5 Robustness, reliability and safeguards

CHAPTER 6 Data Governance and Privacy in AI systems

CHAPTER 7 Emphasizing Transparency in AI Operations

CHAPTER 9 Societal and Environmental Implications of AI systems

CHAPTER 10 Accountability and Redress

1

Exploring the AI Landscape

Few chapters in the grand narrative of technological innovation have been as loaded with twists, turns, and fanfare as the rise of artificial intelligence. What began as a speculative dream of machine cognition has become a central pillar of modern infrastructure, powering tools that shape our work, guide our decisions, and increasingly govern key aspects of daily life. In 2025, AI tools are no longer confined to research or novelty; they are embedded in the core functions of government, business, and society. But this transformation has also forced urgent questions to the forefront: how should AI be governed? Who is accountable when it fails? And can legal systems keep pace with AI systems that evolve faster than institutions? This chapter maps the AI landscape as it stands today, offering both technical and regulatory context for the challenges explored throughout this book.

AI in 2025: From hype to infrastructure

"It will either be the best thing that's ever happened to us, or it will be the worst thing. If we're not careful, it very well may be the last thing," famous scientist Stephen Hawking supposedly said on artificial intelligence. A decade ago, such statements captured the public's ambivalence toward a technology that was still largely theoretical for most people. Popular culture played a major role in shaping early hopes and fears, with stories of sentient machines, killer robots, and benevolent assistants offering a narrative scaffolding for what AI could become. But today, the debate is no longer about what AI might bring: it's about what it has already delivered – from synthetic media that reshapes elections, to predictive models used in policing and healthcare, to generative tools that challenge intellectual property and professional identity.

AI is already here

AI is embedded in the systems we use every day, often without recognizing it as such. From personalized shopping and content feeds to resume screening tools and predictive

By the end of this chapter, you'll be able to ...

- Analyse how AI systems have become embedded in social, economic, and legal infrastructures.
- Identify the key ethical, legal, and governance challenges posed by contemporary AI.
- Trace how AI has shifted from experimental technology to a systemic force shaping public policy.

Four science fiction movies that cemented the concept of AI as large and in charge.

maintenance systems in industry, AI quietly shapes decisions, behaviors, and outcomes across nearly every sector. Much of it doesn't even go by the name "AI": it's packaged as automation, personalization, optimization. And yet, it's AI all the same.

The reason the label "AI" has returned to prominence is not because the technology is new, but because its reach has become impossible to ignore. If we look at the history of the field, we see recurring waves of hype and disillusionment – periods known as the "AI Summers" and "AI Winters." During the winters, developers rebranded their work under less controversial terms like analytics or automation, often to secure funding or avoid inflated expectations.[1] But that pattern broke with the emergence of very capable general-purpose foundation models, which can be integrated in various AI systems. Systems like ChatGPT, Gemini, and Claude brought astonishing capabilities: coding, summarizing, translating, drafting contracts, even composing music and art.

This ubiquity has triggered a dual response. On the one hand, AI is hailed as a breakthrough on the scale of the internet itself, a cornerstone of the so-called Fourth Industrial Revolution and a strategic technology foundational to society as a whole.[2] On the other, it has intensified public concern: about surveillance, about bias and opacity, about job displacement and dependency. Governments, regulators, and civil society are no longer asking whether AI will disrupt society. They are grappling with the fact that it already has.

The rise of algorithms

At the core of every AI system lies the algorithm: a sequence of instructions or rules that guide the system's behaviour. While the term may evoke modern connotations of inscrutable black-box logic, its origins are much older. The word "algorithm" derives from the Latinized name Algorismus, a reference to 9th-century Persian mathematician Muhammad ibn Musa al-Khwarizmi.[3] His work introduced formal methods for calculation and problem-solving, laying the foundations for both algebra and the procedural thinking that now underpins computing. In its simplest form, an algorithm is like a recipe: a clear, step-by-step method to solve a problem or reach a result.

Today's algorithms, however, are rarely simple or transparent. The ones that power AI systems operate at a scale and speed that far exceed human capacity. Unlike earlier forms of automation, which replaced manual labor, AI systems are designed to perform cognitive tasks: evaluating options, recognizing patterns, interpreting language, even making decisions.[4] This is the defining shift. A robotic arm might assemble a product with physical precision, but an AI system determines how to optimize production and assess risk, develop content for marketing campaigns, and sell the product to businesses and consumers.

Why is this happening? Despite the grand narratives about AI transforming humanity, the real driver is far more mundane: efficiency. AI enables organizations to process more data, make more decisions, and execute more tasks, all faster, cheaper, and without the variability of human judgment.[5] It's not just about augmenting human intelligence; in many cases, it's about eliminating the human altogether. From customer service chatbots to automated legal screening and predictive analytics in management, AI replaces slow, expensive, hard-to-control people with scalable code. Where past waves of automation replaced manual labour, AI now targets cognitive work: analysis, evaluation, and decision-making.

New risks from reliance on AI

The growing reliance on AI systems also introduces new structural risks, in particular because these systems replace human decision-making in areas once thought to require context, judgment, or empathy. When optimized for scale and efficiency, AI systems often perform exactly as instructed, yet produce outcomes that are biased, unjust, or harmful. The root issue is not simply malfunction, but misalignment: between what the system optimizes for and what society actually values.

Here are some illustrative cases from the past several years:

1. **Apple Card credit limits (USA, 2019):** Apple's credit algorithm, managed by Goldman Sachs, gave drastically lower credit limits to women than to their male spouses with

shared finances. Despite the company denying gender bias, it could not explain how the algorithm made decisions.

❷ **Dutch SyRI welfare profiling system (Netherlands, 2020):** A government system designed to detect welfare fraud used opaque indicators that disproportionately flagged low-income, immigrant-heavy neighborhoods. A court ultimately ruled it violated human rights due to its discriminatory effect and lack of transparency.

❸ **Ofqual exam grading algorithm (UK, 2020):** During the COVID-19 pandemic, the UK's Office of Qualifications used an algorithm to assign high school grades based on school averages rather than student performance. The system downgraded students from disadvantaged schools, triggering mass protests and a full reversal.

❹ **Predictive policing with HART (UK, 2018–2022):** The Durham Constabulary's Harm Assessment Risk Tool classified suspects as low, medium, or high risk of reoffending. With accuracy barely above chance (around 54%), the system nonetheless influenced real decisions, raising serious concerns about automation bias and profiling.

❺ **DALL·E and Midjourney image generation bias (global, 2022–2025):** Studies have shown that image generators trained on web-scale data disproportionately depict doctors as white men and nurses as women, even when prompted neutrally. Attempts to "correct" this with overcompensation (e.g. diversity prompts) have raised concerns about artificial representation efforts.

Each of these cases highlights a different failure mode: biased training data, lack of explainability, inappropriate proxies, over-reliance by human users, or poorly defined objectives. Yet the common thread is clear: when AI systems are embedded into decision-making, they tend to reflect and reinforce existing inequalities, this time at machine speed and scale. Left unchecked, these failures become normalized as part of "efficient" governance or service delivery.

Another structural risk is transparency, or rather the lack thereof. Many AI systems, especially those based on deep learning, produce outputs without any clear rationale or explanation. This "black box" problem undermines accountability, especially when decisions by AI systems may impact fundamental rights such as access to services, education, or employment. Users and regulators alike face situations where neither the developer nor the deployer can fully explain why a decision was made, only that it was made by a system optimized to perform well on paper. As the Council of Europe observed, "human beings feel they have no control over and do not understand the technical systems that surround them".[6] In such contexts, contesting an AI-driven outcome becomes nearly impossible.

The risks also extend deeply into personal data and surveillance. AI systems often rely on massive data sets scraped from digital platforms, public records, sensors, and user interactions, often collected without clear consent or even awareness. Facial recognition tools can track people in public space; behavioral models can infer sexual orientation, political beliefs, or mental health status from seemingly innocuous metadata. Even anonymized datasets have been shown to be re-identifiable when combined with external sources. This pervasive data extraction erodes individual privacy, flattens boundaries between public and private life, and creates the infrastructure for continuous, invisible surveillance.

Lastly, the rise of generative AI has triggered an intense legal and policy battle over intellectual property. The large foundational models are typically trained on humongous datasets that include copyrighted text, imagery, and audiovisual content, often collected from unauthorised sources without permission or compensation. While EU law allows for data mining under certain conditions, rights holders can opt out, and the legal clarity around AI training uses remains limited (we will address this in Chapter 6). Artists, journalists, and publishers have raised alarms that their work is being used to train systems that now compete with them in the market, generating 'original' outputs derived from their own copyrighted materials. As generative AI tools become more commercially dominant, the conflict between AI development and IP protection is becoming one of the defining regulatory challenges of the decade.

The call for ethics in AI

An underlying concern with the cases we've just discussed is that AI systems are often deployed without a clear ethical grounding. To an extent, this is understandable: new technologies typically require a period of observation and reflection before ethical standards can be meaningfully calibrated. Yet, as is common in the information and communications technology (ICT) sector, the adoption of AI has been so rapid, and its integration so wide-reaching, that this reflection has lagged behind. The result is a mismatch: systems with massive social impact, embedded in everyday decisions, but developed and deployed before core questions of fairness, accountability, and human dignity were even asked – let alone answered.

The 2018 'techlash' as catalyst

The call for ethics in AI moved from niche academic debate to mainstream concern in the late 2010s, accelerated by scandals that made algorithmic power visible to the public.[7] The Cambridge Analytica revelations, algorithmic discrimination in credit and hiring, and Big Tech's role in manipulating public discourse sparked what became known as the "techlash", a debate about power, accountability, and the role of AI in shaping democratic life.

While many companies responded with ethical codes and principles, these efforts often lacked mechanisms for enforcement or follow-through. Ethics became branding ("virtue signaling").[8] The gap between promise and practice quickly became clear: without legal obligations, few organizations made meaningful changes, because *"[w]hen ethical ideals are at odds with a company's bottom line, they are met with resistance".*[9]

The shift to regulation

This realization laid the groundwork for a shift from voluntary ethics to rights-based regulation. Policymakers, especially in Europe, began grounding AI governance in fundamental rights: non-discrimination, privacy, (human) autonomy, and transparency. A landmark initiative was the AI4People framework (2018), which formulated five overarching principles of ethical AI we still see today:[10]

- Beneficence: Promoting well-being, preserving dignity, and sustaining the planet
- Non-maleficence: Privacy, security and "capability caution"
- Autonomy: The power to decide (whether to decide)
- Justice: Promoting prosperity and preserving solidarity
- Explicability: Enabling the other principles through intelligibility and accountability

These principles were picked up by the EU's High-Level Expert Group on AI (2019) and UNESCO's global recommendations (2021), which would later form the ethical backbone of binding instruments like the AI Act, and continue to inform how we assess AI systems today.

To regulate, govern, or even assess AI in any meaningful way, we first need to understand what it is and is not. That turns out to be harder than expected. Artificial Intelligence remains a shifting concept, shaped as much by cultural expectations as by technical capacity. In the next section, we map the spectrum of AI systems to help clarify what needs governance, what carries risk, and where current legal definitions draw the line.

Navigating the AI spectrum

To govern AI effectively, we first need to understand what we're governing. That task remains surprisingly elusive. Artificial Intelligence is not a single technology, but a shifting label applied to a spectrum of systems; some rule-based, some adaptive, some capable of self-optimization. It covers everything from facial recognition and voice assistants to deepfake generators and large language models. Before we can assess legal risk or ethical impact, we must ask a deceptively simple question: what counts as AI?

The evolution of AI

The term artificial intelligence was famously introduced in 1955 as *"the problem of making a machine behave in ways that would be called intelligent if a human were so behaving."*[11] That definition, though charmingly indirect, already hinted at a core ambiguity. It doesn't claim machines are intelligent, but only that their behaviour can resemble intelligence in a human.

Early AI systems took the form of what we now call expert systems. These systems attempted to formalize human expertise into "if-then" logic trees: if the temperature is above 38°C and there's a cough, then conclude "fever". This symbolic approach was deterministic and transparent. But it was also brittle. Any situation not foreseen by the rule set could not be handled, and maintaining or scaling the knowledge base quickly became unmanageable.

Modern AI, by contrast, is dominated by machine learning (ML). Instead of coding the rules directly, ML systems are trained on data. They infer patterns from past examples and use statistical models to generalize. A medical ML system doesn't "know" what fever is in symbolic terms. Rather, it learns from thousands of cases that certain combinations of symptoms tend to co-occur. The shift from expert systems to ML marked a profound transition: from systems we could inspect and edit, to systems that emerge from data, often in ways we don't fully understand.

Crucially, many legal and policy frameworks still implicitly picture the early type of AI when regulating the modern form. The idea that an AI "reasons" through interpretable steps underpins many calls for explainability, traceability, and meaningful human oversight, and thus creates confusion and friction when applied to systems that do not, in fact, reason at all, but derive outcomes statistically from patterns in data.

Machines that think

Imagine: you are in a room, receiving a series of Chinese symbols through a slot in the door. Using a comprehensive rulebook, you methodically shuffle and reorder these symbols, then send out the reorganized symbols through another slot. To an external observer who understands Chinese, it appears as if you're proficiently answering questions in the language. But here's the catch: you don't understand a word of Chinese. You're merely processing symbols according to rules. This is the essence of John Searle's "Chinese Room" thought experiment. If you, as the symbol processor, don't truly understand Chinese, can we claim that a machine, which processes information based on programmed algorithms, genuinely "understands" or "thinks"?

As noted above, The question of whether AI "thinks" has haunted the field since its inception.[12] Systems that produce fluent language, detect patterns, or make recommendations are often assumed to "understand" what they're doing. But as the thought experiment shows, symbol manipulation is not the same as comprehension.

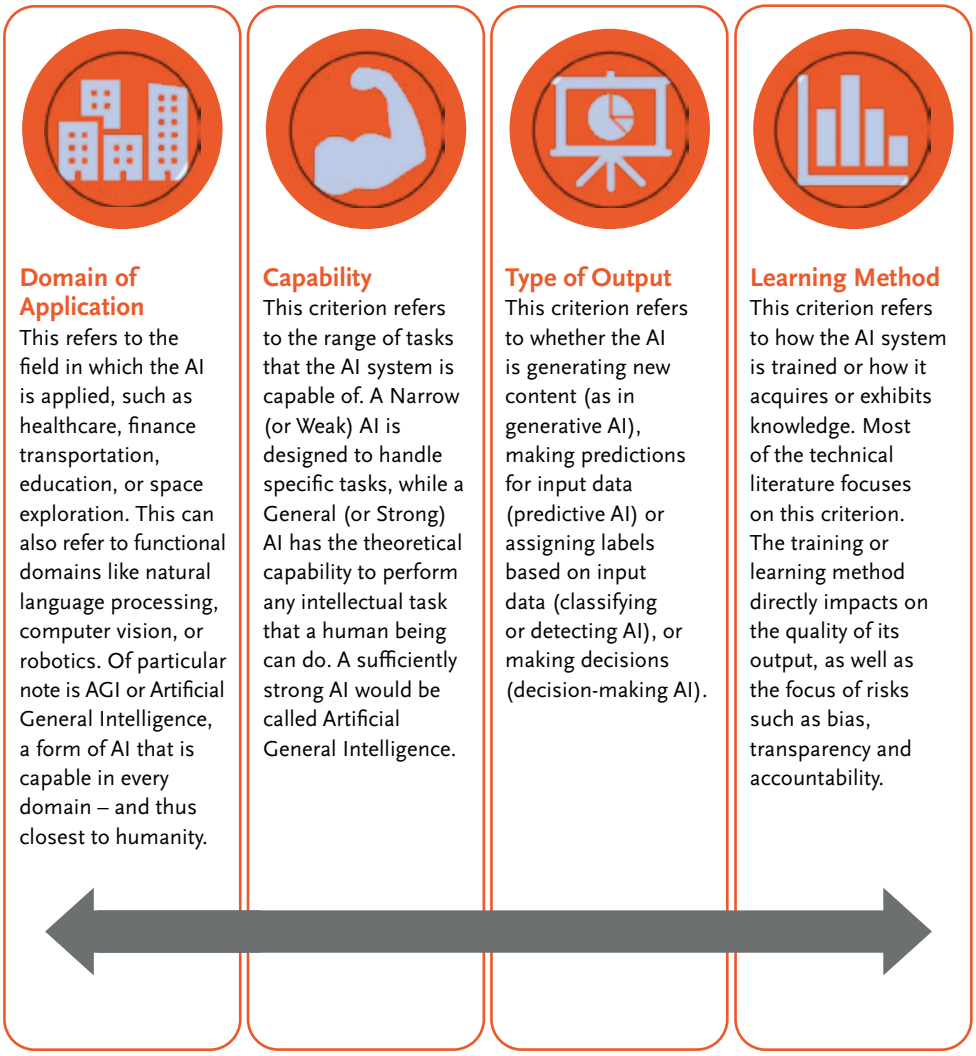
Today's machine learning systems, especially large language models like GPT or Claude, are the most powerful instantiation of this effect. They can summarize legal documents, write poems, or draft code with striking coherence. But they do not "reason" in the way humans do. They approximate reasoning behaviour by predicting plausible continuations of language based on vast statistical correlations. This gives the appearance of intelligence, but without grounding in meaning, intent, or awareness. These systems can answer questions about ethics or law without having any concept of right, wrong, or justice. In other words, they are not thinking – they are simulating thinking.

Recent advancements have made this distinction even harder to spot. Large language models now exhibit so-called "long-form" or "deep" reasoning: the ability to carry out multi-step operations, maintain consistent positions over extended interactions, and decompose complex prompts into intermediate tasks. They can produce arguments, legal memos, and output intermediate thoughts and conclusions along the way. Using tool-augmented reasoning (where models call external functions or access calculators, databases, or code interpreters) these become even more cognitively capable than ever. However, what's often mistaken for reasoning is in fact post-hoc justification. Large language models don't "think through" a problem to arrive at an answer. They generate responses based on the statistical likelihood of word sequences, often producing the conclusion first and then assembling a seemingly coherent rationale around it.[13] This phenomenon of "post hoc rationalisation" is particularly visible in cases where the model makes an error but explains it convincingly, or flips positions mid-dialogue without acknowledging the shift. The result is not logical inference, but rhetorical coherence, plausible reasoning that gives the appearance of deliberation without the underlying process. This poses serious challenges for regulatory frameworks that assume explanations are causal and traceable, rather than reverse-engineered narratives constructed on the fly.

And yet, the idea that AI might one day surpass human intelligence – the so-called singularity – continues to capture public imagination and shape governance anxiety.[14] Whether seen as a breakthrough or a threat, the singularity implies not just smarter machines, but systems that exceed human control, comprehension, or alignment. For now, that remains speculative. But the real concern is more immediate: that we are already outsourcing critical human functions to machines that appear to think, while still lacking intent, understanding, and accountability.

A taxonomy of AI

Artificial Intelligence (AI) is an overarching term that encapsulates a wide variety of subfields and technologies. Many have attempted to provide a taxonomy of AI.[15] It is impossible to provide one hierarchical structure that covers all forms of AI, but there are several main criteria according to which an AI system could be compared against others, as illustrated below.



In the context of this book, we find the 'learning method' criterion most suitable. This choice aligns well with the learning objective of enabling the application of legal and ethical norms in AI practice, with data governance and transparency and accountability on the other. When organizing AI according to learning method, the main division is

between expert systems and machine learning. Zooming in on machine learning, we can make the following subdivision:

- **Supervised learning** trains a model on labeled examples: a dataset where each input (e.g. an image, sentence, or transaction) is paired with the correct output (e.g. a label, category, or value). The model learns to generalize from these pairs to make predictions on new, unseen inputs. This is the backbone of many commercial AI systems today, from spam filters and medical diagnosis tools to credit scoring and facial recognition.
    - Supervised systems depend heavily on training data quality. Key risks here are bias in the training data and lack of representativeness, which can lead to discriminatory outputs. Auditing data provenance and annotation processes is essential.
- In **unsupervised learning**, the system receives input data without any labels or predefined outputs. The goal is to find patterns, groupings, or structures within the data. This is commonly used for clustering customers, detecting anomalies, or compressing large datasets. Two main forms are Latent Variable Models (LVMs), often used to reduce complexity or discover hidden themes, and Generative Adversarial Networks (GANs), which learn to generate new data by pitting two networks against each other until one (the discriminator network) can no longer identify the other's (the generator's) output as synthetic.
    - These systems can surface unexpected or non-intuitive groupings, which can influence decision-making without clear justification. Their opacity demands rigorous validation and monitoring.
- **Self-supervised learning** leverages unlabeled datasets by automatically generating labels ("pseudo-labels") from the data itself. For example, in large language models (LLMs), words in a sentence may be masked, and the model attempts to predict these missing words based on surrounding context. Self-supervised learning is the dominant learning method behind current foundational AI models like GPT-4, Gemini, or image generation models such as DALL·E. Self-supervised learning differs from unsupervised learning as it explicitly creates pseudo-labels from data itself, while unsupervised methods find latent structures or patterns without explicit labeling.
    - Self-supervised learning models inherit biases and gaps from the enormous, often opaque, datasets they are trained on. Intellectual property rights (IPR) are a significant governance concern, given the massive scale of publicly available, but potentially protected, data scraped during training.
- **Reinforcement Learning** involves agents that learn by interacting with an environment and receiving rewards or penalties based on their actions. It's used in areas like robotics, game-playing, real-time bidding in ads, and increasingly in dialogue agents and autonomous systems. A recent advancement, Reinforcement Learning from Human Feedback (RLHF), is now commonly used in training LLMs. It helps align

model outputs with human preferences by incorporating human evaluations into the reward signal.

- Reinforcement Learning systems can exhibit unpredictable behavior, especially when reward functions are poorly defined or when they exploit unintended shortcuts. Oversight must focus on goal alignment and testing under varied conditions.

➔ **Transfer Learning** is the concept of adapting a pre-trained model using a new, often smaller, dataset. For example, a language model trained on internet-scale text might be fine-tuned for legal document summarization, or an image model trained on generic photos might be adapted to detect tumors in radiology scans. Fine-tuning can be as simple as prompt tuning (modifying how output is to be presented) or as complex as a full retraining of the model, but typically involves a balanced approach using techniques like LoRA (Low-Rank Adaptation) to update only a small part of themodel.

- Using transfer learning, downstream systems inherit risks and biases from their base models, which are often trained on opaque data. Fine-tuned models may behave unpredictably in edge cases or when assumptions from the pretraining phase no longer hold.

➔ The term **Deep Learning** refers to a family of methods (often applied in the above categories) based on neural networks with many layers. It enables high performance in image recognition, natural language processing, and speech synthesis. Deep learning is what powers most modern AI breakthroughs. A specific class of deep learning are Convolutional Neural Networks (CNNs), particularly powerful for processing images.

- Deep models are notoriously hard to interpret. Their scale increases risks related to explainability, adversarial attacks, and energy consumption. Their opacity also complicates impact assessments and human oversight.

➔ A new phenomenon, emerging in the 2020s, are **foundation models**. These are trained on massive datasets using unsupervised or self-supervised learning and can be fine-tuned for specific tasks. They can perform multiple functions, such as translation, summarization, classification, even reasoning within a single architecture. Examples include large language models (e.g., GPT-4, Claude), multimodal systems (e.g., Google Gemini), and image generators (e.g., DALL·E, Midjourney). In the EU's AI Act these are regulated under the name of general-purpose AI (GPAI) models (see Chapter 3). Note that in a legislative context, the term general-purpose AI is preferred; "foundational model" is common in a technical context.

- GPAI training data is often opaque, their outputs highly persuasive but error-prone, and their applications so broad that defining responsibility becomes difficult.

➔ Most modern foundation models are built using **transformers**, a neural architecture that uses self-attention mechanisms to process sequential data (like text, code, or

images). In this context, 'attention' refers to a mechanism that allows the model to dynamically focus on the most relevant parts of the input (e.g. specific words in a sentence or regions in an image) when generating each output. Transformers have largely replaced recurrent and convolutional networks in state-of-the-art AI.

- Transformers make systems harder to interpret or constrain. Their ability to process and generate text, code, and imagery across domains blurs boundaries between tasks.

From models to agents

The evolution of learning methods and architectures has culminated in an emerging class of AI systems often described as agentic. These systems do not merely predict or classify. Rather, they plan, act, and adapt across sequences of tasks. Sometimes referred to as AI agents, they represent a shift from reactive models to systems that exhibit autonomous, goal-directed behaviour. However, it is important to note that these systems do not truly reason. They simulate planning by chaining statistically plausible steps together, often guided by predefined tools or prompts. What emerges may look like deliberation, but it lacks understanding, intent, or internal logic.

Agentic systems are typically built on top of foundation models, enhanced with planning mechanisms, memory, tool use, and in some cases the ability to self-decompose complex objectives into smaller steps. Frameworks such as AutoGPT, LangChain, and AutoGen have made it easier to orchestrate these capabilities into interactive workflows. For example, an agent tasked with "analyze this contract and generate a risk summary" might search legal sources, write draft text, check for consistency, and refine its response without further user intervention.

What makes agentic AI distinct is not its architecture per se, but its behaviour: systems that decide what to do next rather than waiting for instructions. They may generate, observe, re-evaluate, and reattempt tasks in response to intermediate feedback. Some can even invoke other agents to solve subtasks, leading to emergent patterns of behaviour. Agentic AI complicates oversight, testing, and responsibility, particularly when outcomes emerge from multiple steps taken across tools and time.

Defining Artificial Intelligence

Artificial intelligence is one of the most widely used and poorly defined terms in the digital age. Across policy papers, technical reports, and marketing decks, it refers simultaneously to logic-based rule engines, adaptive machine learning systems, and generative models that write poetry or simulate voices. A common theme, though, is that AI generally is whatever machines can't *quite* do. Once a task becomes routine

(like route planning, voice recognition, or credit scoring) it tends to lose its AI label and becomes just another background function.[16]

Behaviour as criterion

When Alan Turing proposed what is now known as the Turing Test for AI, he wasn't attempting to define intelligence in abstract terms.[17] The common way the outcome is told, is that if the interrogator cannot distinguish between human and AI after a sufficiently long conversation, the AI can be regarded as intelligent. However, what Turing actually said was that the question "Can machines think?" was "too meaningless to deserve discussion" and used the experiment to illustrate that the distinction cannot be meaningfully made. The famous Dutch computer scientist Edsger Dijkstra put it succinctly as: *"The question of whether a computer can think is no more interesting than the question of whether a submarine can swim."*[18]

In doing so, Turing introduced a pragmatic shift that continues to shape how we approach AI today: intelligence should be judged not by how a system works internally, but by how it behaves externally. It is a perspective that still underpins many popular and legal understandings of AI today. If a chatbot gives coherent answers, if a recommendation engine feels intuitive, if an image generator produces something visually compelling, then we often treat those systems as intelligent, regardless of how they work.

This behavioral framing paved the way for the concept of AI agents: systems that perceive their environment, make decisions, and act upon that environment to achieve goals. While the idea of intelligent agents dates back to the 1990s and early 2000s, it has gained renewed prominence in the 2020s with the emergence of autonomous assistants, planning-capable language models, and task-oriented agentic systems like AutoGPT and Devin. Unlike traditional chatbots or classifiers, these systems combine reasoning, memory, and action, often chaining together multiple steps across tools and contexts. The agent metaphor makes the behavioral criterion explicit: the system is judged by what it does, not whether it "understands."

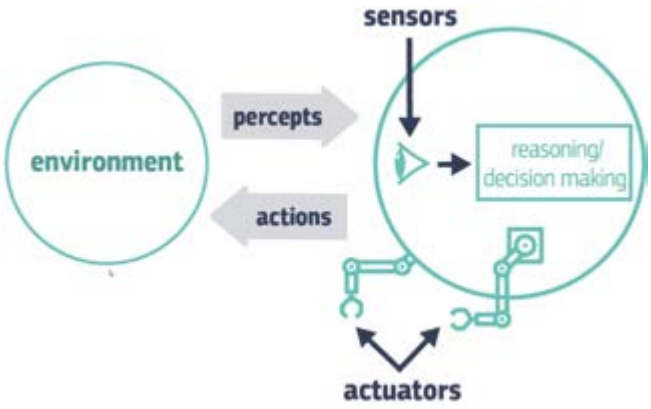
Autonomy as a definitional key

In their 2018 Communication that kicked off the process that would ultimately produce the AI Act, the European Commission gave a loose definition of AI: *"Artificial intelligence (AI) refers to systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals."* While this definition still uses the problematic term 'intelligence', it did provide a hint for a new approach to define AI: autonomy.

Autonomy, in this context, does not mean self-awareness or volition. It means that the system can operate without direct, real-time human instruction, responding to inputs and generating outputs that affect its environment. Whether selecting which content to recommend, adjusting pricing dynamically, or executing a multi-step plan to book travel, an autonomous system acts with a degree of discretion. This capacity to act without being told what to do next is what makes AI distinct from traditional software and what gives rise to legal, ethical, and social concerns.

Autonomy thus becomes a definitional anchor for governance. It provides a functional criterion for inclusion under legal frameworks: not how smart a system is, but whether it can observe, decide, and act with minimal human oversight. In practice, this also means that AI is defined less by what it is, and more by what it does, and the risks those actions pose.

The Commission's loose definition was formalized by its subquently founded High-Level Expert Group (HLEG) in the starting document *Definition of Artificial Intelligence*. As shown below, it describes AI systems as having the ability to perceive their environment through sensors, interpret this data to make decisions, and act upon the environment through actuators.



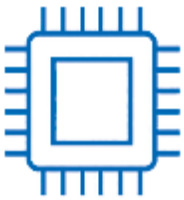
The concept of AI autonomy is closely linked to the fundamental rights of human dignity and freedom (Articles 1 and 6 of the Charter): when we delegate more decision-making power to AI, we consequently reduce the level of control that humans exert, which can potentially infringe upon our autonomy and dignity. This concept is also reflected in the 2016 GDPR, that in article 22 provides that no person shall "be subject to a decision based solely on automated processing". We will examine the concept of autonomy in Chapter 4.

The legal definition: AI in the AI Act

The European Union's AI Act builds on this autonomy-oriented understanding. In Article 3(1), it defines an AI system as *"A machine-based system designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments."*

This definition is notable for what it includes, and what it avoids. It does not require intelligence, consciousness, or human-like behaviour. Instead, it emphasizes:

1. Machine-based implementation (software or hardware);
2. Autonomy (and optionally adaptiveness);
3. Objectives to strive for;
4. Inference, as opposed to human rule-following;
5. Outputs with real-world influence (both in the physical and virtual world).



Machine

A machine can be as simple as a computer executing a model. What matters is that they are machine-operated, not human-executed.

Autonomy

This criterion ranges from partially automated tools to fully autonomous agents.

Objectives

The system must act according to explicit or implicit objectives, such as 95% accurate forecasting, a high satisfaction score or operating within guardrails.

Inference

The core distinction between AI and traditional software: AI infers outputs from input or training data, rather than following fixed, human-coded rules.

Influence

Can be both physical and virtual. A smart lawnmower qualifies as much as a music recommender algorithm that starts new tracks, or an HR system that generates employee assessments into the file.

This approach reflects the choice for risk-based governance that underpins the AI Act. The more autonomously and adaptively a system operates (and thus the more impact it has on society) the more scrutiny it demands. In practice, this means that even simple systems may count as AI under the law, if they operate with some of autonomy, and are capable of acting with inference and consequence.

General-Purpose AI: A new regulatory category

One of the most significant additions in the final version of the AI Act is the formal introduction of general-purpose AI (GPAI). This responds directly to the rise of large, flexible foundation models like GPT-4, Claude, Gemini and the open-source LLaMA model released by Meta. These are not built for a single application, but can be adapted to countless downstream uses across sectors. They are referred to as 'models', distinguishing them from the 'systems' just discussed. Most generally, models are not machine-based and cannot function independently without supporting elements, e.g. hardware, user interface, and data pipelines. Therefore, they do not have the capability to influence their environment.

The AI Act defines GPAI as an AI model that "displays significant generality and is capable of competently performing a wide range of distinct tasks regardless of the way the model is placed on the market and that can be integrated into a variety of downstream systems or applications". This definition stresses that GPAI models are not narrowly designed for a single task or domain, and more importantly that they are intended for others to build upon. These models challenge regulation because their risks are not tied to a specific deployment. The same AI model might be used in a harmless writing assistant, a critical medical interface, or a high-stakes law enforcement tool.

To add to the confusion, GPAI models can be deployed as AI systems, e.g. in the form of web-based tools, APIs, or integrated software layers. For example, prompting ChatGPT through a browser may seem like simply using a third-party service, but in regulatory terms, it can qualify as deploying an AI system, particularly if its outputs inform decisions in a high-risk context. The line becomes even less clear when a developer builds a custom GPT using OpenAI's tools: is that just an interface variation, a new system, or something in between? And what if a company uploads its own domain-specific data, writes a 100-line instruction prompt, and uses that model to screen job applicants or generate legal guidance? At what point does customization become development? At what point does use become deployment?

In the next two chapters, we will examine how these broad definitions feed into a risk-based regulatory structure, and how organizations must classify and assess the AI systems and GPAI models they develop, trade, use, or deploy. But before we do that, let's have a quick look at emerging trends.

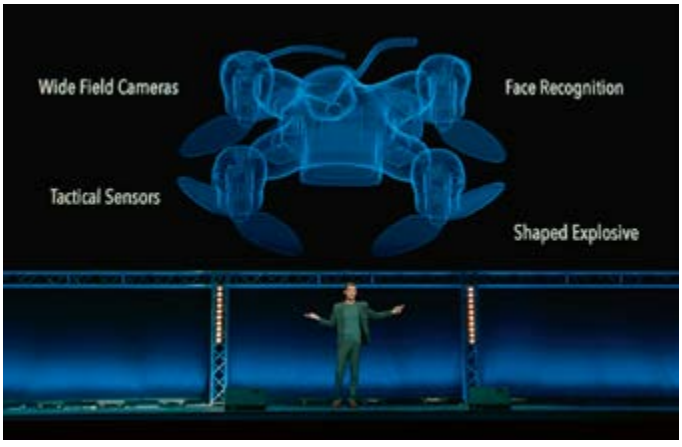
Looking ahead: Emerging trends

As we look ahead to 2025 and beyond, artificial intelligence is no longer defined by its novelty or potential, but by its integration into critical domains of power, infrastructure, and sovereignty. The technologies once treated as experimental are now embedded in military operations, healthcare systems, intellectual property conflicts, and even orbital space governance. These developments signal a shift from hypothetical ethics to concrete governance: not what AI might do, but what it is already doing – and whether our institutions can still shape its trajectory. This section outlines four domains where AI is rapidly transforming global systems: warfare, healthcare, intellectual property, and space – each raising distinct challenges for law, accountability, and long-term strategic control.

AI and military systems

The war in Ukraine has made a reality what military analysts have long predicted.[19] AI is now used in target recognition, drone swarming, battlefield logistics, cyber operations, and ISR (intelligence, surveillance, reconnaissance). Autonomous and semi-autonomous drones have been deployed for both defensive and offensive missions, sometimes making real-time decisions in contested airspace. This has put the debate on Autonomous Weapons Systems (AWS) in a new light.[20]

Legally, this domain remains underregulated. The EU AI Act explicitly excludes military AI from its scope (Article 2 (3) and Recital 24), leaving such systems outside the risk-based framework that governs civilian uses. There are no international laws or treaties to oversee or limit AWS specifically, although international humanitarian law (such as article 36 of Additional Protocol I to the Geneva Conventions) does provide that any new weapons system must be assessed for lawfulness prior to being deployed on the battlefield. This assessment includes a rigorous assessment of predictability in its



A fictional autonomous weapon concept from the 2017 short film 'Slaughterbots', illustrating concerns about potential misuse of AI and drone technology.

operation and its reliability in the field. One can imagine such an assessment for the SGR-A1, but for the weaponized drone swarm from the 2017 arms-control advocacy video Slaughterbots this will not be possible.

On December 2, 2024, the United Nations (UN) General Assembly adopted a resolution on Lethal AWS with overwhelming support, proposing a two-tiered approach to prohibit some while regulating others under international law. The timeline for a binding treaty formalizing these intentions still seem far away.

AI and healthcare

Healthcare is one of the most promising and contested frontiers of AI. From diagnostic support and personalized medicine to surgical robotics and mental health chatbots, AI is increasingly woven into clinical workflows and patient-facing systems. In high-resource settings, AI is used to interpret radiology scans, triage emergency patients, and detect patterns in electronic health records. In low-resource contexts, AI chatbots are used to supplement mental health services, support health literacy, or predict local disease outbreaks where infrastructure is lacking.

But this integration comes with legal and ethical complexity. Unlike decision-support systems of the past, many modern AI tools are capable of adaptive, autonomous decision-making: choosing which symptoms to highlight, which treatment protocols to suggest, or when to escalate a case. This introduces risks not just of technical failure, but of systemic harm, especially when AI-generated outputs are either overtrusted or treated as deterministic by busy healthcare professionals. A particularly urgent concern is the phenomenon of hallucination, where large language models generate information that is factually incorrect, fabricated, or misleading – yet presented in fluent and authoritative language. In a clinical context, hallucinations can lead to the suggestion of non-existent drugs, fictitious case citations, or dangerously inaccurate interpretations of test results. Combined with persistent bias in training data and the absence of well-defined fallback mechanisms, these systems can amplify rather than mitigate risk.

Despite these uncertainties, the performance of AI in healthcare is rapidly advancing. Systematic reviews and meta-analyses now confirm that AI systems – particularly in imaging fields like dermatology, ophthalmology, and radiology – can match or exceed human specialists in some diagnostic tasks. A 2023 meta-review in The Lancet Digital Health found that deep learning models achieved comparable accuracy to clinicians across 14 specialties.[21] In 2025 the FUTURE-AI Consortium proposed consensus guidelines including a set of 30 best practices were defined, addressing technical, clinical, socioethical, and legal dimensions.[22]

A notable emerging trend in AI and healthcare is the rise of carebots: AI-powered systems designed to assist with caregiving tasks in clinical, home, or elder care settings.[23] These systems can monitor vital signs, issue medication reminders, and even engage in conversation to reduce loneliness. While they offer potential relief in understaffed healthcare environments, they also raise pressing ethical concerns.[24] Carebots collect large volumes of sensitive health data, triggering questions about data protection, informed consent, and surveillance in private spaces. Moreover, by assuming roles traditionally filled by human caregivers, they risk depersonalizing care, replacing human empathy with simulated affect.

In the European context, AI in healthcare is regulated not only under the AI Act, but also through the Medical Device Regulation (MDR, Regulation 2017/745), which governs software used for medical purposes. Many AI-based systems, such as diagnostic support tools or treatment planning software, qualify under MDR and thus must undergo conformity assessment before entering the market. This creates some regulatory overlap, as we will see later in this book.

Intellectual property in AI creations

"Intellectual property shall be protected", as article 17 of the Charter of Fundamental Rights of the EU puts it. The goal of protecting creative works such as books, music, movies or sculptures, and protecting technological innovations has long been a part of society. But today this simple phrase sits at the center of one of the most complex regulatory dilemmas in the AI landscape: how to reconcile copyright protection with the rise of generative AI systems trained on massive, often unlicensed datasets.

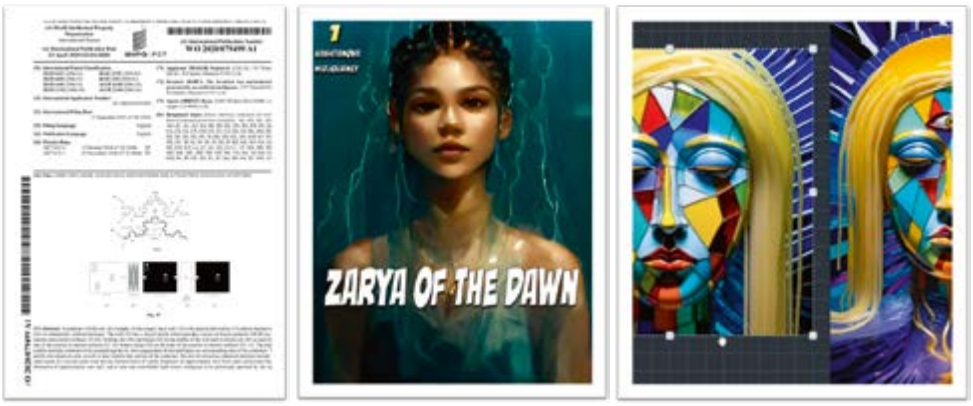
Large-scale foundation models such as GPT-4, Claude, Stable Diffusion, or Midjourney are trained on vast corpora of web-scraped content – including books, articles, images, and source code, almost all of which is copyright-protected. These systems do not simply "learn" general knowledge; they absorb statistical patterns from original works and reproduce them in outputs that may closely resemble, or even partially reconstruct, their source material. While on the one hand this is clearly a reuse of protected material, on the other hand this type of use is utterly incomparable with traditional copying, appropriation or piracy. It is literally unprecedented.

The legal landscape is still forming. In the United States, several high-profile class action lawsuits (e.g., Andersen v. Stability AI, Silverman v. OpenAI) are testing whether AI-generated outputs constitute derivative works, and whether training models on copyrighted data violates the rights of authors and creators. The outcomes of these cases remain pending, but they are likely to shape global precedents.[25]

In the European Union, the regulatory approach is more structured but no less contentious. Under the 2019 Copyright in the Digital Single Market Directive, AI developers can mine text and data for scientific purposes or general use – but only if rights holders do not opt out (article 4). An open issue is that opt-out signaling must occur "in a machine-readable fashion", which requires technical standards that are as yet not available.[26] We will revisit this subject in more detail in Chapter 6 on data governance.

At the same time, generative AI is forcing a fundamental reevaluation of what constitutes creativity and authorship. Traditionally, intellectual property law has required a human author for copyright protection to apply. But AI-generated content now blurs the line between tool and creator. A prominent example is the DABUS case, where AI researcher Stephen Thaler sought to register patents naming an AI system as the inventor (leftmost figure below). While rejected in most jurisdictions, DABUS triggered global legal debate and set a precedent for questioning the human requirement in invention.[27]

In the copyright realm, Kris Kashtanova's graphic novel Zarya of the Dawn, partially generated using Midjourney, in 2023 received copyright registration in the U.S., only to be revoked once the AI's role was revealed (middle figure). The US Copyright Office appears to have changed its position in 2025 in the case of "A Single Piece of American Cheese" (rightmost figure). Here, a human creator used an AI image generation tool to repeatedly adjust AI output, creating what copyright law calls a "collector's copyright", creativity being found in the human "selection, coordination, or arrangement" of the elements in the work.



AI and embedded political power

A growing concern of the rise of AI is its growing role in undermining democratic processes. Electioneering is increasingly shaped by algorithmic targeting, automated propaganda, and synthetic media. While digital manipulation is not new, the scale, personalization, and realism introduced by generative AI have fundamentally changed the threat landscape.

Generative models now produce deepfake videos, synthetic voices, and AI-written campaign messages that are nearly indistinguishable from authentic political content. These tools have been used to impersonate candidates, spread false endorsements, and suppress voter turnout through misleading messages. In recent election cycles, including the U.S., EU Parliament, and several African and Southeast Asian states, AI-generated content was deployed in microtargeted disinformation campaigns with minimal attribution or oversight.

Compounding this is the use of LLMs in political messaging, voter persuasion, and even direct voter interaction. Chatbots trained to simulate a candidate's tone or platform are already in use, raising urgent questions about authenticity, transparency, and consent in digital campaigning. Unlike traditional media, these systems can tailor their language and logic to each user, creating a fragmented and manipulable information space.
Democratic institutions are struggling to respond. Transparency mandates for political advertising have not kept pace with synthetic content. Election regulators often lack the technical capacity to identify or trace generative material. And while the EU's Digital Services Act (DSA) and the AI Act contain obligations for transparency and traceability, enforcement remains piecemeal and post hoc. The rise of "algorithmic intermediaries" that mediate political participation itself is deeply concerning.[28]

AI legislation for space exploration

As the final frontier of this chapter, let's consider deployment of AI in space activities, offering solutions to challenges in data collection, space traffic management, and space debris removal. Prominent examples include the KubeSat satellite management platform and the Artificial Intelligence Learning Earth Observation (AILO). Famously, the NASA Perseverance Mars Rover used AI technology to find its way on the Red Planet.

However, international space law, in particular, lack adequate regulation of this sphere of activity fraught with serious risks.[29] The existing major treaties on outer space adopted in the 60s and 70s, for obvious reasons, do not contain provisions governing the use of artificial intelligence technologies and establish only general principles covering any activity in outer space and on celestial bodies. One example is the 1972 Liability Convention on International Liability for Damage Caused by Space Objects,

that today regulates liability for damage caused by a space object. The very generic rules in this Convention could be applied to damage caused by AI, e.g. a miscalculated course causing a crash or other accident. However, the establishment of fault presupposes the existence of due care standards, which is quite problematic in the case of using such new technologies.[30]

As with other fields of technology, the integration of AI in space activities is transforming the way we approach the field. This underlines the need for comprehensive and up-to-date regulations, which in turn shows why the AI Act in Europe is such a monumental step. It's clear that as we move further into the era of AI-driven space exploration, law and policy must keep pace with technological advancement.

Moving on, let's consolidate the main points from this discussion in our final section, "Key Takeaways", to ensure we've captured the breadth and depth of AI's impact on our legal and ethical landscape.

Key takeaways

In this opening chapter, we unraveled the cultural and political aura surrounding artificial intelligence. We traced the shift from speculative concern to concrete governance, grounded in the real-world consequences of algorithmic decision-making and the urgent demand for embedded ethics. We examined how AI is no longer a standalone technology but a systemic force, increasingly integrated into the machinery of healthcare, political communication, military operations, creative industries, and even orbital space. These domains reveal AI's power to shape not only outcomes, but the institutions and values that govern our societies. Questions about hallucinations, authorship, democratic integrity, and off-world accountability are no longer science fiction – they are governance problems.

With this foundational perspective on AI's definitions, dilemmas, and frontiers, we now turn to the legal core of this book: the European Union's AI Act. In the next chapter, we examine how Europe has attempted to draw boundaries, assign responsibilities, and turn ethical ambition into regulatory architecture.

2

The AI Act: Structure, scope, and impact

In this chapter, we examine the European Union's Artificial Intelligence Act (AI Act), the world's first comprehensive legal framework for AI. We will explore why the EU saw the need for dedicated AI legislation, how the Act defines its scope, and the risk-based logic that structures its core obligations. This includes the distinction between prohibited, high-risk, and limited-risk systems, as well as the emerging legal treatment of general-purpose AI. Questions of implementation and internal compliance – how to actually follow these rules – will be the focus of Chapter 3. Let's begin with the question: why did Europe decide that AI needed its own legislation – and how did it evolve from ethical principles to enforceable law?

Understanding Europe: Trustworthy AI defined

The European drive to regulate AI was sparked by the ethical concerns prompted by the quick rise of AI in the past years.[1] It is however also part of a broader approach, called the Digital Decade 2030. This approach seeks to transform and digitize the European economy and society. It is part of the so-called European Green Deal, which seeks to transform the EU into a modern, resource-efficient and competitive economy.[2] Both were initiated after the 2008 worldwide crisis. As a result, there has been intense activity in the EU concerning the regulation of digital markets, including the AI sector.

Trustworthy AI as a key principle

Within this vision, artificial intelligence is seen as more than just an economic driver. It is a governance challenge, a technology that must be shaped by democratic principles, human rights, and public accountability from the outset.[3] That framing was first formalized in 2018, when the European Commission convened the High-Level Expert Group on Artificial Intelligence (HLEG), a multi-stakeholder advisory body tasked with drafting guidelines for what Europe calls *trustworthy AI* – AI that deserves our trust.

By the end of this chapter, you'll be able to ...

- Understand the structure, scope, and purpose of the EU AI Act
- Analyze the Act's risk-based classification system and its legal implications for different types of AI systems.
- Identify the core legal obligations for high-risk and general-purpose AI under the Act's framework.

The HLEG's 2019 *Ethics Guidelines for Trustworthy AI* formulated a three-pronged approach. Trustworthy AI must be lawful, ethical, and robust.[4] The ethical aspect was further developed in those guidelines, using seven key aspects that we will explore in detail in chapters 4 through 10. The lawful aspect of trustworthy AI was codified in the AI Act, adopted 12 July 2024. The Act in turn sets basic requirements for technical and societal robustness, including cybersecurity (see chapter 5), which are to be further worked out in so-called harmonised standards that are expected to be set in 2026 and 2027.

To help operationalize these principles, the HLEG in 2019 developed the *Assessment List for Trustworthy AI* or ALTAI, a self-assessment tool for developers and deployers with a wide range of questions. You'll see these appear throughout chapters 4 to 10 to help you operationalise the content in your work. Many of ALTAI's elements foreshadowed later legal requirements in the AI Act.

Towards the AI Act

The Guidelines were part of the EU's AI strategy announced in 2018.[5] In this strategy, existing legislation on topics like product liability, ecommerce, motor vehicles would be updated to address applications or impact of AI, and a new AI Act would complement them to with general rules based on the risk level of the AI technology. An explicit part of this strategy was that the ethical values inherent in such new rules should be exported to other countries in the world. While this may sound ambitious, it is not unjustified: the so-called "Brussels Effect" of EU regulation having worldwide impact is a proven phenomenon in international politics and economics.[6] The 2016 General Data Protection Regulation (GDPR) is also widely cited as exhibiting a similar effect in other countries.

A first draft of the AI Act was released by the European Commission on 21 April 2021. While the legislative process, negotiations and lobbying was intense[7], no one foresaw what would become the single greatest disruption to the Act's design. In late 2022, OpenAI released ChatGPT, a conversational interface to a large-scale foundation model trained on web-scale data. Within months, it gained over 100 million users – faster than any software in history – and triggered a wave of public attention, enterprise adoption, and governmental anxiety. What had been a niche technical debate about foundation models suddenly became front-page news. These models could write essays, summarize legal texts, generate policy briefs, and simulate human conversation at scale. Yet, they were not tied to a specific high-risk use case and therefore fell through the legal cracks of the original AI Act proposal.

In a rush to respond to political and public pressure, lawmakers introduced an entirely new section to the Act during late-stage negotiations: a dedicated chapter on General-Purpose AI or GPAI models. The chapter sets a dual strategy. First, providers of GPAI models are subjected to a baseline set of obligations: transparency, documentation, and usage policy requirements. Second, a stricter regime applies GPAI models having so-called *systemic risk,* risks with the potential to disrupt or influence broad sectors of society. These models must implement specific risk mitigation measures to prevent or reduce large-scale harms.

The concept of systemic risk is borrowed from financial regulation, where it first gained prominence after the 2008 global financial crisis. In the digital policy domain, the term was later adopted in the Digital Services Act (DSA) to describe Very Large Online Platforms (VLOPs) like Facebook, TikTok, and Instagram—entities whose algorithmic reach grants them the power to shape public discourse and influence democratic processes. By applying this concept to GPAI, the AI Act acknowledges that certain AI models now operate at a societal scale, requiring regulatory tools previously reserved for infrastructure-like actors.

The result was a compromise: a law initially designed to regulate purpose-specific AI had to stretch mid-process to accommodate infrastructure-scale models that defied earlier categories. While hailed as a regulatory first, it also revealed the tension at the heart of AI governance: lawmaking moves slowly, but AI development and adoptiondoes not.

The AI Act provides a staggered introduction. First, in February 2025 the requirements for AI literacy and the prohibitions against certain unwanted AI uses entered into force. From August 2025 onwards, GPAI providers must comply with basic requirements and supervisory authorities must have been established. In August 2026, the rules for high-risk use cases (see next section) must be observed, followed a year later by a compliance requirement for producers of AI as safety components in regulated products.

Understanding the AI Act

With 113 legal clauses, 180 introductory recitals providing context and intent and 13 detailed annexes, the AI Act rivals the GDPR in complexity and interpretational challenges. Let's examine the structure and key terminology of the AI Act in more detail.

A table of contents

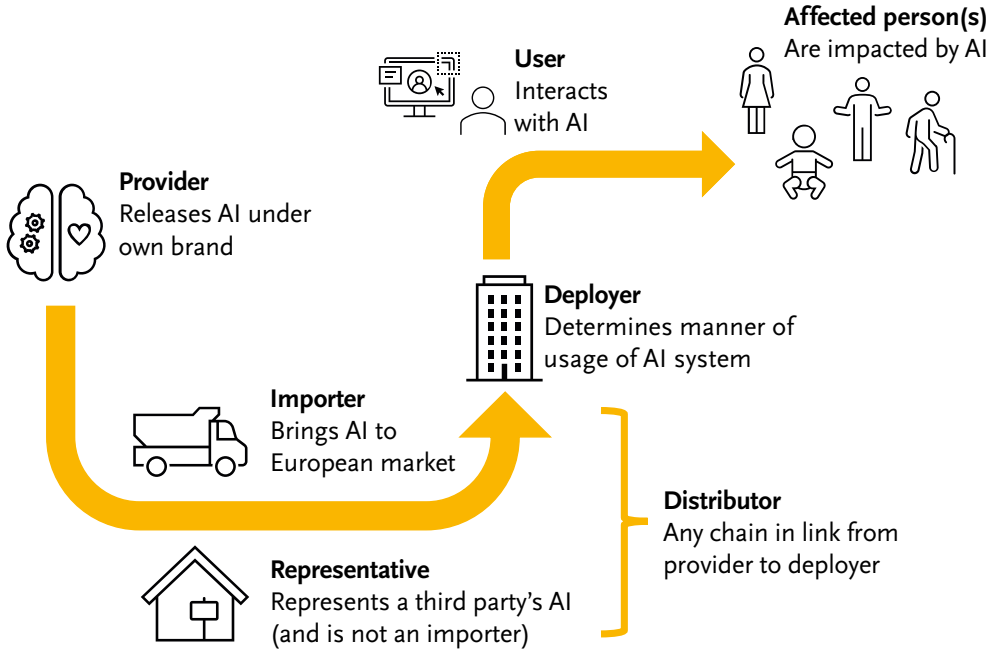
The AI Act counts 113 articles, divided into thirteen chapters, each of which may contain several sections.

1. Chapter I – General Provisions. The first articles define subject matter and scope of the AI Act, including key definitions (discussed in the next subsection) and general tasks for national legislators and supervisory authorities.
2. Chapter II – Prohibited AI Practices. This chapter prohibits certain applications of AI because they fundamentally contradict core values of the EU, more on which in the next section on managing risk.
3. Chapter III – High-risk AI Systems. In this chapter, divided into four sections, the AI Act sets out the many compliance requirements and restrictions for so-called "high-risk" AI. These are the subject of the following chapters of this book.
4. Chapter IV – Transparency obligations. This chapter defines basic transparency obligations for any type of AI, not just high-risk systems.
5. Chapter V – General-Purpose AI. Measures to regulate the large foundation models were introduced partway through the AI Act's adoption in response to the meteoric rise of ChatGPT.
6. Chapter VI –Measures in support of innovation. The regulatory sandbox is the main focus of this chapter. We'll meet the sandbox later in this chapter.
7. Chapter VII – Governance. This chapter sets up the governance and enforcement structure in the EU, including the AI Office that coordinates activity of the national supervisory authorities throughout the Union.
8. Chapter VIII – EU Database for High-Risk AI systems. To further stimulate transparency, this chapter establishes a publicly-accessible database listing every high-risk AI system deployed in the European Union.
9. Chapter IX – Post-market monitoring and market surveillance. In this chapter various rules regarding information sharing and market monitoring are established. Of particular note are a duty to report 'serious incidents' regarding the use of AI systems. This chapter also establishes the investigative powers and competence of the supervisory authorities.
10. Chapter X – Codes of Conduct and Guidelines. This chapter establishes a mechanism for self-regulation of AI systems that do not qualify as high risk. These could set requirements for example on environmental sustainability, accessibility, stakeholders participation and diversity of development teams. It also permits the Commission to issue Guidelines to interpret and clarify aspects of the AI Act, such as the definition of "AI system" or the application of requirements for high-risk AI providers.
11. Chapter XI – Delegation of Power. This formal chapter provides the legal basis for the European Commission to unilaterally amend certain parts of the AI Act, notably the list of use cases considered high-risk.
12. Chapter XII – Penalties. This chapter provides the legal limits for financial penalties for violations of the Act.

⓭ Chapter XIII – Final Provisions. The last chapter of the AI Act amends other legislation to refer to the AI Act, sets its date of entry into force and introduces certain exemptions for AI systems already on the market.

Key terminology

The AI Act contains over 50 definitions to ensure consistent application of its provisions. The key terms have been illustrated in the below chart, which visually shows the growth of an AI system from creation to market operation.



- **AI system** – A machine-based system that infers outputs (e.g. predictions, recommendations, decisions) from input data, operating with some degree of autonomy, and capable of affecting physical or virtual environments.
- **Provider** – The natural or legal person who develops an AI system or has it developed and places it on the market under their name or trademark.
- **Deployer** – The entity using an AI system under its authority (e.g. employers, government agencies, hospitals).
- **User** – The person interacting with an AI system, but not legally responsible for its deployment.
- **Importer** – The entity that brings an AI system developed outside the EU into the EU market.

- **Distributor** – An actor in the supply chain who makes the AI system available without altering it.
- Together, the provider, the deployer, the authorised representative, the importer and the distributor are called the **operator(s)** of an AI system.
- **Affected persons** are those persons or groups who are subject to or otherwise affected by an AI system. This is not necessarily the same as 'users', as one may be affected by an AI system's action without actively using it – or even being aware that this is happening. For example, if a self-driving car operates on the open road, any traffic participant would be an affected person, while only the driver in the car could be called the 'user' of the system

Each entity in what the Act calls the "AI value chain" is assigned different obligations, which we'll tackle in more detail in chapter 3. This chapter will also examine the compliance implications of customizing, embedding, or deploying GPAI systems in real-world use cases. For now, it's important to understand that the legal classification depends not just on what the model is, but on how it is used, modified, and placed on the market.

The definition that stands out the most is that of an **AI system**. According to Article 3(1) of the AI Act:

> *"'AI system' means a machine-based system designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments."*

The key operative word here is **"infers."** In legal interpretation, this separates AI from traditional software. Essentially, the question is who made the rules that the system operates under? Traditionally human programmers did that job, but AI systems are unique in that they derive or infer rules by analysing data looking for patterns (compare chapter 1's list of machine learning techniques).

At the basic level, inference includes

- Identifying patterns or anomalies in data (e.g. fraud detection, machine malfunction predictions)
- Generating language or content (e.g. large language models, image generators)
- Recommending actions (e.g. treatment prioritization, route optimization)
- Classifying or clustering data dynamically (e.g. customer segmentation, document tagging)

Automation and workflows based on static, deterministic rules (e.g. "if A then B") would be excluded.

The distinction unfortunately was blurred somewhat by the European Commission's *Guidelines on the definition of an AI system*, which attempt to exclude standard uses of statistical techniques that also underlie most AI systems. The clearest interpretation is to distinguish between on the one hand statistical methods used to assist human decision-making and machine learning systems that set, adjust, and act on rules autonomously. For example, a regression model helping a doctor choose a threshold for intervention is not AI under the Act; but a system that trains on patient data to automatically set that threshold and trigger clinical alerts without human input would qualify.

The AI Act's definition of an AI system includes the possibility that a system may "exhibit adaptiveness after deployment". Most machine learning systems do not adapt once deployed; they operate based on a static model trained before release. If performance declines or data changes, these systems typically require retraining by developers. However, some systems can adjust parameters or update behavior while in use, e.g. a robot vacuum cleaner that learns over time which floor surfaces cause more resistance and adjusts the speed or angle of its brushes accordingly. This post-deployment adaptiveness is optional under the definition; a system can qualify as AI even if it remains static after deployment, as long as it performs inference from data autonomously.

Under the definition, an AI system must be capable of "influencing physical or virtual environments." For physical systems, this is relatively straightforward: a robot that navigates a room, adjusts machinery, or administers treatment to a patient clearly acts on the physical world. For virtual systems, the interpretation is more vague. The Guidelines take a broad reading: an AI system that produces outputs consumed by other software systems can already be said to influence a virtual environment, such as a Web service that returns a risk score fed directly into an HR platform to filter job applicants or a content moderation system whose outputs determine whether posts are flagged or removed on a social media platform. The same can be said for AI agents, that autonomously carry out actions in the virtual world.

Material and geographical scope

With the proper terminology established, we can discuss the material and geographical applicability of the AI Act. Material applicability means if a system qualifies as an AI system, while geographical applicability means whether an action with such an AI system by a particular entity triggers the AI Act's obligations.

For **material** applicability, first of all the system must meet the definition of an AI system quoted above. But there's more: the AI system must have been 'put into service', which roughly means any supply of the system for distribution or use on the Union market in the course of a commercial activity, whether in return for payment or free of charge. The intent behind this convoluted sentence is to exempt research and non-commercial activities such as open-source development of publicly available AI systems.

For **geographical** applicability, the AI system must have been placed on the market or put into service in a European Union member state (or Iceland, Liechtenstein and Norway, the European Economic Area members not part of the EU). "Placing on the market" is the EU term for making a product available for sale in a member state. "Putting into service" is the corresponding term for services. It can refer to both offering to third parties (e.g. as an internet service like a chatbot) and using internally, like a knowledge base available to employees.

There is a broader option: if "the output produced by the system is used in the Union", the AI provider or deployer behind such output is subject to the AI Act. This option thus does not require specific sales or other commercial activities in a member state. If a European firm were to hire, say, a Canadian provider to use an AI system to evaluate the effectiveness of its marketing activities, the output of that system would be used in the Union and thus the Canadian provider would have to comply with the AI Act.

It is irrelevant whether the provider placing the AI system on the market or putting it into service is established within the Union or in a third country. The same goes for providers, importers and distributors. The AI Act thus has a broad reach: a US-based provider that offers access to an AI system through a website is required to comply with the AI Act, despite not having a physical presence in the Union. In practice, the provider would need to appoint an authorised representative who would assume these responsibilities (and face the administrative penalties, including fines, if any violation occurred).

Entry into force and transitional provisions

The AI Act entered into force on July 12, 2024 with a two-year transitional period during which AI providers, distributors and deployers will be able to adjust their products, services and processes to the new requirements. However, the prohibited practices (see next section) and AI literacy obligations already are applicable from February 2, 2025 onwards. Market surveillance authorities that oversee the AI Act gained their powers to investigate and sanction violations by August 2, 2025, as did the European Commission's powers to oversee the large GPAI providers like OpenAI, Microsoft, Google and Anthropic. As of August 2, 2026, high-risk applications must meet a series of requirements.

The transitional provisions are complex. High-risk systems that are already on the market on August 2, 2026, do not have to comply with the law as long as they are not "significantly" modified. However, the definition of "significant" is very unclear, partly because they rely on the conformity assessment – which does not have to be done for certain systems.

Managing risk: AI practices and their classification

The AI Act takes a risk-based approach: rather than specifically prohibiting or regulating certain AI systems or use cases, the Act defines three levels of risk and attaches legal obligations and limitations to each. To understand the implications, let's first take a step back at what type of risk we are referring to.

AI and fundamental rights

The Charter of Fundamental Rights of the European Union lays the foundation for the EU's legal system. The terms 'risk' and 'harm' in the context of AI regulation must be understood as a reference to these fundamental rights and values of the Union. Deployment of an AI system may introduce a great variety of harms. Here are a few examples of commonly-cited harms, written in terms of fundamental rights with reference to the relevant provisions in the Charter.

- **Privacy and Data Protection (Articles 7 & 8):** AI's capability for mass surveillance and data processing could lead to invasions of privacy, conflicting with the right to the protection of personal data and the respect for private and family life.
- **Non-Discrimination (Article 21):** Biased AI algorithms could result in discriminatory outcomes in services, employment, and justice, which would contravene the principle of non-discrimination.
- **Freedom of Expression (Article 11):** Overzealous AI moderation tools could restrict lawful speech, impinging upon the right to freedom of expression and information.
- **Workers' Rights (Article 31):** AI in the workplace could lead to intrusive surveillance and job displacement, undermining the right to fair and just working conditions.
- **Right to a Fair Trial (Article 47):** AI tools used in legal proceedings could lack accountability and transparency, threatening the right to an effective remedy and to a fair trial.
- **Consumer Protection (Article 38):** AI that deceives or manipulates consumers, or that fails to ensure product safety, could violate the right to a high level of consumer protection.

- **Protection of Personal Integrity (Article 3):** AI applications in medicine or biometrics that misuse personal health data or bodily information would conflict with the right to integrity of the person.
- **Environmental Protection and Sustainability (Article 37):** AI systems, through their lifecycle from development to deployment and disposal, can have significant environmental impacts. The energy consumption required for training complex AI models and the electronic waste generated from rapid obsolescence of AI-enabled devices can contribute to environmental degradation.
- **Freedom to Conduct a Business (Article 16):** AI systems could potentially disrupt markets by enabling monopolistic behaviours or by creating unfair competitive advantages through data dominance or algorithmic collusion. This could lead to significant economic harm for smaller companies that cannot compete with AI-enhanced businesses.

There is still significant discussion among scholars how to weigh these harms.[8] However, the AI Act does not require providers or deployers to make individual risk assessments for each AI system. The "risk-based" aspect of the AI Act refers to the fact that the law itself provides that judgment. (Some deployers must perform a so-called fundamental rights impact assessment or FRIA, which we'll discuss in chapter 10.)

Three levels of risk

The AI Act has as its primary aim the mitigation of risks, which it defines as the combination of the probability of an occurrence of harm and the severity of that harm. There is a three-tier approach to managing the aforementioned risks:

1. **Prohibited Practices.** These provide manipulative, exploitative and social control practices that contradict the fundamental rights, and thus should be banned entirely. Prohibited practices are listed in Article 5 of the AI Act, which means adding or removing such a practice requires a revision of the Act itself.
2. **High-Risk Practices.** While not contradicting fundamental rights or Union values, these practices pose significant risks of adversely impacting these rights. AI systems exhibiting such a level of risk must implement a large number of compliance requirements before being permitted on the market. High-risk use cases are listed in Annex I and III of the AI Act, which the European Commission can amend without revising the Act itself.
3. **Transparency Issues.** AI systems both high-risk and not may come with transparency issues. The first issue is whether users can recognize an AI as such. More importantly is whether *output* of an AI system is recognizable as such. (Some literature refers to this issue as the "low risk" or "minimal risk" category, but the AI Act does not have this term.) We will discuss this issue in chapter 7.

Prohibited practices

The AI Act prohibits a series of AI-driven practices as being particularly abhorrent: use of AI for "manipulative, exploitative and social control practices" that contradicts Union values of respect for human dignity, freedom, equality, democracy and the rule of law and Union fundamental rights, including the right to non-discrimination, data protection and privacy and the rights of the child.

The list is in article 5:

❶ **Subliminal manipulation that distorts behavior**, such as a retail AI that modulates in-store lighting or scent to subconsciously steer vulnerable customers toward high-risk financial products.

❷ **Exploitation of vulnerabilities**, such as an educational chatbot that targets children with persuasive messages to spend money in an online learning app.

❸ **Social scoring by public authorities**, such as a municipality ranking citizens based on tax behavior, social media activity, and healthcare compliance, then denying certain permits based on the score.

❹ **Real-time remote biometric identification in public spaces for law enforcement (with narrow exceptions)**, such as when deploying live facial recognition at a protest to identify and track individuals.

❺ **Predictive policing based on profiling or location**, such as a city using an algorithm to assign higher police presence to neighborhoods based on residents' socio-economic profiles and historical arrest rates.

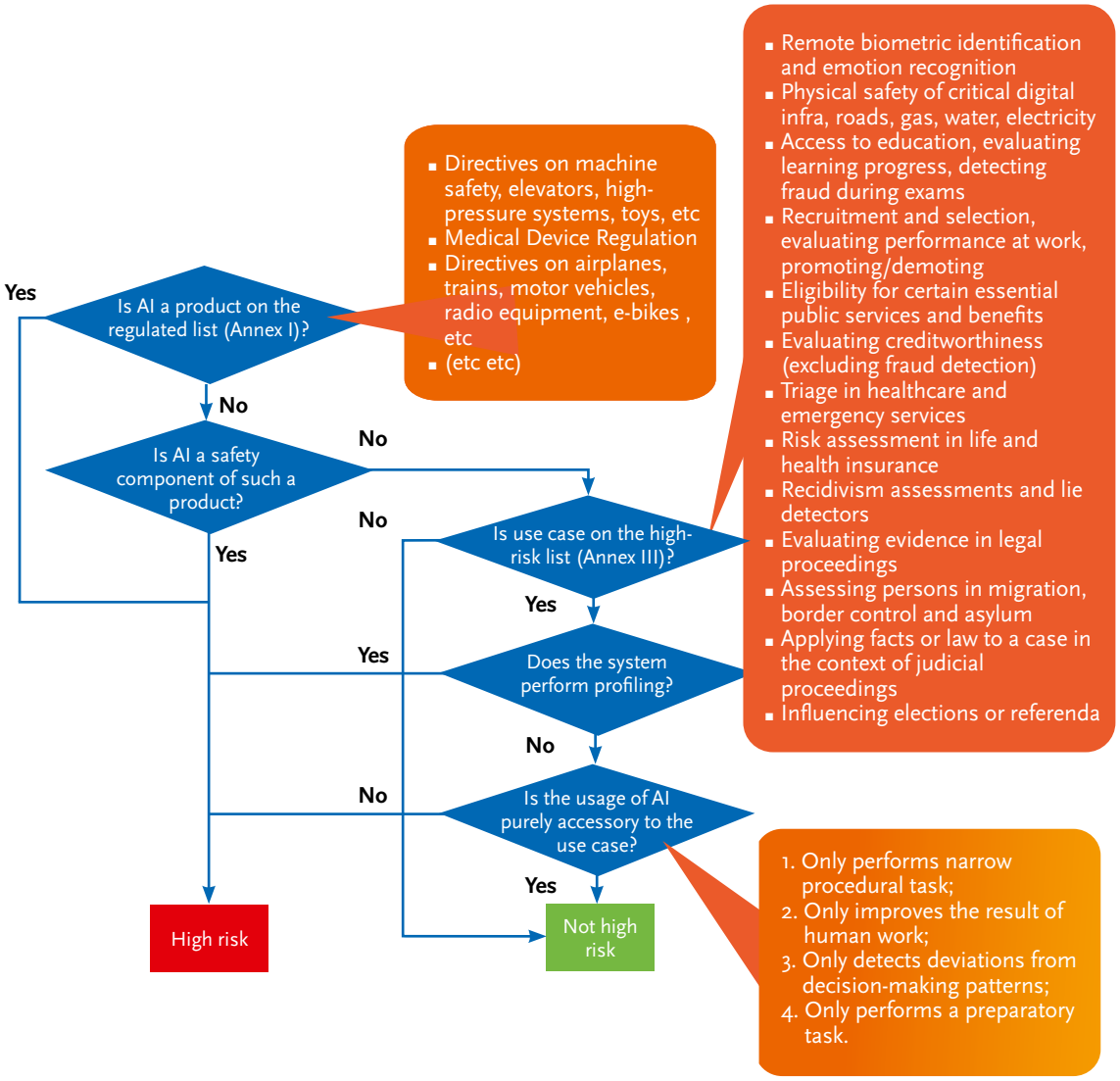
❻ **Emotion recognition in the workplace and education**, such as a proctoring tool in online exams that flags students as suspicious if it detects facial expressions labeled as "nervous" or "unengaged." Note that while the prohibition speaks of "workplace", it also applies to job applicants.

❼ **Biometric analysis to infer protected categories**, such as ethnic background, political opinions, trade union membership, religious or philosophical beliefs, sex life or sexual orientation.

There is no way to claim an exception for a prohibited practice, not even with the informed consent of all persons involved. The ban is absolute and there are no exceptions. The only real way to avoid the prohibition is to adjust the practice so that it avoids the thresholds of article 5.

Determining high-risk AI

For the determination process of high-risk AI there is more flexibility. This process is illustrated in the flowchart below.



The term 'high-risk' thus is the outcome of a formal process. The first item to check is whether the AI system qualifies as a regulated item under a variety of EU product legislation listed in Annex I. This includes regulations and directives such as the safety of production machinery, elevators or toys, the regulation of radio equipment, gas ovens and notably medical devices (MDR). If an AI system is intended to be used as a safety component of such a regulated product, or is itself such a product, the AI system is by definition high-risk.

If no listed EU regulations apply, the process moves on to Annex III which lists a number of use cases and applications of AI. They are divided into eight sections with generic titles, but those are not decisive: only if the AI system matches the description of a given use case does it qualify as "high-risk". For instance, section 2 speaks of "critical infrastructure" but defines only "safety components in the management and operation of critical digital infrastructure, road traffic, or in the supply of water, gas, heating or electricity" as high-risk. An AI system managing safety of a city's Metro rail infrastructure for instance thus is not high-risk, simply because it is not covered by this definition.

Some AI systems are used within a high-risk use case, but by themselves do not introduce a significant risk of harm. For example, an AI system may merely provide guidance to persons applying for welfare, without materially influencing the outcome of the decision on that application. Such AI is excluded as being "purely accessory". However, in order to be allowed to invoke this exception, the provider must make a written assessment prior to deployment of the AI system. A fine can in theory be provided for misclassification.

While a provider is free to provide any argument, the AI Act lists four use cases that by definition are sufficient to fall within the exception:

1. The AI performs only a narrow procedural task, such as automatically adjusting the brightness of streetlights based on environmental conditions.
2. The AI only detects deviations from decision-making patterns, e.g. by highlighting an outcome that is unusual for the inputs given.
3. The AI does not influence actual decision-making, but merely provides inputs such as predictions of crowd behaviour or labelling of documents provided by users.
4. The AI only improves the quality of work, e.g. acting as a copy editor or by identifying likely areas for cleanup by sanitary workers.

Finally, when an AI system involves profiling of natural persons (the use of personal data to evaluate certain personal aspects, as defined by the GDPR), a provider may not invoke the exception. For instance, take an AI system used in a public employment service that highlights which job seekers are statistically less likely to find work – based on personal data such as age, education level, or migration background – only flags these cases for further human review. While this may appear to be deviation detection, the system is performing profiling under the GDPR and would therefore not be eligible for the exemption.

General-Purpose AI and Foundation Models

The most consequential addition to the AI Act during the legislative process was the introduction of a new category: **General-Purpose AI (GPAI)**, sometimes also called **foundation models**.[9] These are AI systems trained on broad datasets for a wide range of tasks, often not tied to a specific application at the time of development. Their defining feature is generality: the same underlying model may be used in legal advice, education, customer service, creative writing, or even medical triage, depending on how it is deployed.

GPAI models are subject to a two-tiered regulatory regime, depending on their capabilities. All GPAI providers have to create technical documentation (e.g. training data summaries, model design rationale), instructions for use and usage policies, and quality testing results. Models that have so-called 'systemic risk' must additionally assess and mitigate these risks, maintain detailed model documentation and capabilities evaluations, ensure an adequate level of cybersecurity protection and report serious incidents (see chapter 1).

What Is General-Purpose AI?

A general-purpose AI model is one that "displays significant generality and is capable of competently performing a wide range of distinct tasks". It is intended to perform generally applicable functions, such as image and speech recognition, audio and video generation or question answering. The assumption is that such models are primarily intended for integration into downstream systems and services, where they are either used as-is (e.g. via API access) or adapted for specific applications.

However, in practice, we increasingly see GPAI models deployed directly to end users, e.g. as chatbots, writing assistants, content generation tools, and more. The AI Act accounts for this by distinguishing between the GPAI *model* and the GPAI *system*. When a foundation model is used in a consumer-facing product, it must comply with both the GPAI obligations and the general rules for AI systems, including risk classification and transparency.

The two-tiered GPAI framework

The AI Act imposes a two-tiered regime for GPAI, recognizing that while all foundation models pose certain baseline risks, some may reach a scale or level of impact that requires enhanced obligations.

All GPAI providers must comply with a set of foundational requirements. These include:

1. Technical documentation, including the model's architecture, training methods, evaluation benchmarks, and known limitations.

❷ Integration documentation, allowing so-called downstream providers to incorporate the GPAI model in their products.
❸ Summary of training data, giving meaningful transparency about the type and origin of training data (e.g. text corpora, scraped web content, annotated images).
❹ Usage policies and instructions – Clear terms of use must be provided, along with safeguards to mitigate foreseeable misuse in downstream applications.

These requirements reflect the fact that GPAI models, even when not deployed in high-risk domains, can be repurposed or combined in ways that introduce new legal and ethical concerns.

The AI Act also introduces a second, more stringent tier for GPAI models deemed to carry systemic risk. That is, risk capable of disrupting or influencing broad societal domains, markets, or critical infrastructure. The legal criterion to identify such models is their computational power, expressed as the number of floating-point operations required to train them. During the training phase of an AI model, especially when using deep learning techniques, a massive number of floating-point calculations are performed to adjust the model's learnable parameters. These are one of the most complex calculations for a computer system. This makes the cumulative amount of floating-point operations (FLOPs) needed for a complete training a useful approximation for model capabilities.[10] They are also hard to manipulate, e.g. by using faster hardware.

The Act sets a threshold at 10^{25} FLOPS (that's 10 followed by 25 zeros) of total training compute. The 2020 original GPT-3 was estimated to have used roughly 3×10^{23} FLOPs for its training. To put this in context: a 2025 business laptop running at full computational load (ignoring heat, power, memory, etc.) would need over 3,100 years to perform this number of operations. By contrast, a state-of-the-art AI datacenter, such as one of the EU-funded AI Factories under the Digital Europe Programme, may deliver sustained performance on the order of 10^{20} FLOPs per second across thousands of GPUs running in parallel. At that rate, 10^{23} FLOPs takes just 1,000 seconds (~17 minutes) and the AI Act's threshold of 10^{25} FLOPs would take about 28 hours.

(As we'll discuss in chapter 9, the environmental impact of such a model is tremendous. The carbon footprint of training a model near the 10^{25} FLOPs range can reach hundreds of tons of CO_2, depending on energy sources. Water usage for cooling in such datacenters is also significant. A 2023 study estimated that training GPT-3 consumed over 700,000 liters of water—enough to fill several backyard swimming pools.)

Providers of such models must conduct model evaluations, take measures to identify potential misuse, assess and document systemic risks, and implement risk mitigation strategies covering both technical and societal dimensions. They are also required to

ensure an adequate level of cybersecurity protection and keep track of serious incidents that may occur. At the time of writing, no public GPAI model is confirmed to have reached this scale.

This enhanced regime reflects growing concern over the role of GPAI in generating misinformation, exacerbating inequalities, or creating structural dependencies. The term systemic risk, borrowed from financial regulation and adapted in the Digital Services Act for Very Large Online Platforms (VLOPs), now applies to AI models as infrastructure, and demands a governance response of similar scale.

From GPAI deployer to provider

GPAI is designed to be integrated, but one of the most complex aspects of the AI Act's GPAI regime is determining when a downstream actor ceases to be a deployer and instead becomes a provider of a modified AI system. This distinction matters greatly: as discussed in the next chapter, providers bear the core legal obligations for high-risk AI, including documentation, risk management and conformity assessment.

In Chapter 1, we introduced the concept of transfer learning: the adaptation of a pre-trained general-purpose AI to new tasks or domains. This can range from extensive prompting (where model behavior is shaped through carefully crafted instructions) to fine-tuning using small or large domain-specific datasets. Some adaptations merely adjust how the model presents or structures its output; others involve retraining the model to perform entirely new functions.

When a GPAI model is significantly modified or reoriented toward a new intended purpose, the actor making those changes may no longer be just a deployer, but become a provider of a new AI system. The AI Act however does not set explicit boundaries to make a clear distinction, referring merely to a change in intended purpose but without guidance how extensive this must be. Upcoming guidelines suggest an approach based on the amount of FLOPs needed for the fine-tuning, setting a threshold at 10^{23} FLOPs. Exceeding this threshold triggers a presumption that the downstream actor is now a provider, not merely a deployer. This presumption is rebuttable, but it shifts the burden of proof onto the modifying party, who must demonstrate that their adaptation did not create a materially new model with a distinct intended purpose.

This has major consequences for documentation, risk management, and conformity assessment. As explored in Chapter 3, organizations that integrate or adapt GPAI for sensitive applications must be prepared to meet provider-level obligations, even if they didn't develop the underlying model. The flexibility of GPAI enables innovation – but it also expands the zone of regulatory responsibility.

Addressing innovation: regulatory sandboxes

Regulation of new technology poses a dilemma: regulate too early and risk stifling a valuable innovation – but regulate too late and be left with powerless laws against technology giants that have taken over society. The concept of regulatory sandboxes provides a third way: innovators can experiment with new technologies outside existing regulation, while society as a whole remains shielded from any negative impacts.

Origins of sandboxes

Originally developed in the Fintech sector, regulatory sandboxes create a testbed for a selected number of innovative projects, by waiving otherwise applicable rules, guiding compliance, or customizing enforcement.[11] Typically their scope is limited to experiments, as opposed to large-scale production deployments. Despite this limitation, they have long been criticized as being contrary to key principles of law such as legal certainty, proportionality, and equal treatment. The underlying argument is that those that can operate in the sandbox, can achieve a competitive advantage not available to those who operate in the traditional environment, which may seem unfair.

The fear that AI innovation would be stifled was significant, especially given that most innovations in this space already originate from outside the European Union. This is reflected in the stated goals of AI sandboxes: not only should they facilitate training and testing by AI providers, but also provide supervisory authorities with new insights and the ability to draw up guidance for compliance outside of the sandbox.

AI sandboxes in practice

The AI Act does not itself establish sandbox regimes but rather leaves it to the national supervisory authorities to create appropriate boundaries within European and national law. The modalities and the conditions for the establishment and operation of the AI regulatory sandboxes are to be set down in later legislation, as the AI Act puts it.

Regulatory sandboxes can take many forms depending on the sector, the authority, and the risk level of the innovation. Examples from AI, fintech, and other digital domains include:

- **One-on-one workshops with regulators**, where a startup demonstrates a prototype system and receives informal feedback on likely regulatory triggers (common in early-stage fintech sandboxes).
- **Time-limited test deployments**, where a product is rolled out to a limited user base under pre-agreed safeguards and reporting duties (e.g. FCA fintech sandbox in the UK).

- **Supervised data access pilots**, where innovators are allowed to process real-world personal or health data under strict confidentiality and GDPR guidance (e.g. Norway's AI sandbox, see below).
- **Cross-sector collaboration sandboxes**, where a regulator coordinates multiple stakeholders (tech developers, hospitals, insurers) to test a system in a simulated operational environment.
- **Pre-clearance of novel business models**, where innovators explore how an unconventional service fits into existing law, often leading to informal policy notes or guidance (used by Singaporean regulators).
- **Public-private co-design spaces**, where regulators and developers iterate together on potential AI use cases—sometimes even shaping future technical standards.

One of the most instructive national examples is the AI sandbox operated by the Norwegian Data Protection Authority (Datatilsynet). Launched in 2020, the sandbox offers selected organizations – often startups or public bodies – a chance to work closely with regulators on AI projects that involve complex personal data processing. Participation does not involve exemption from legal obligations. Rather, it allows teams to explore privacy-by-design, risk mitigation, and governance structures in a supervised and supportive environment. Projects have included AI tools for early dementia detection, predictive child welfare interventions, and recruitment algorithms. Crucially, the Norwegian sandbox operates as a collaborative compliance lab: the goal is not to loosen regulation, but to help organizations interpret and apply existing rules responsibly in high-stakes contexts. Its success has made it a model frequently cited in EU-level discussions about scaling regulatory sandboxes across member states under the AI Act.

Related legislation

The AI Act is neither the first, nor the last legislation to address artificial intelligence. It is however the sole European act that specifically regulates this innovative technology. Other laws mainly address outcomes or impact or indirectly regulate behaviour that may be exhibited by AI systems.

The General Data Protection Regulation

Adopted in 2016, the GDPR is the EU's flagship regulation on personal data, which is a fundamental right in the European legal system. Its focus is on lawful, transparent and fair handling of personal data, which is a much broader topic than just AI systems. However, automated decision-making or profiling of persons has been a key point of attention even before the GDPR: already in the 1990s it was a well-established principle that computers should not exclude people or put them at a disadvantage merely by

means of data analysis. The GDPR gave hefty teeth to this principle, with tens of millions in fines available for offenders.

The AI Act and the GDPR obviously overlap where an AI system is trained on personal data or where such a system is used to profile or make decisions. These topics are handled in chapter 6 (data governance) in more detail. In short, the GDPR takes precedence according to the AI Act. An AI system can also affect persons even when no personal data is processed or when the processing is 'anonymous' in GDPR parlance. In such a case, the AI Act would take precedence.

Liability: Product safety and civil redress

Unlike the GDPR, which next to public supervision also allows for individual complaints, the AI Act is primarily enforced through public supervision by national authorities. While authorities can issue fines, restrict usage, or ban non-compliant AI systems, the Act does not provide a direct mechanism for individual citizens to seek compensation.

To address this gap, the European Commission introduced two complementary instruments: a revision to the Product Liability Directive (PLD), the EU's longstanding product safety regime, and a standalone AI Liability Directive. Unfortunately, the latter was abandoned in early 2025, citing lack of a common vision among political actors.

The revised PLD has been adopted and is in its two-year implementation period. Its general thesis is that individuals harmed by an AI system may sue any operator in the value chain. When the individual demonstrates the harm was likely caused by a defect in the product, the operator faces a reversed burden of proof is reversed: it must provide evidence that the product met relevant safety regulations or the defect could not have been prevented given the state of scientific and technical knowledge at the time. This includes harms caused by AI components, although the PLD limits this to harms in the form of death, injury and loss of data.

Harms caused by AI systems may also be addressed through collective claims. Under the Representative Actions Directive, consumer protection organizations and other qualified entities can bring forward collective redress actions on behalf of individuals harmed by unfair, misleading, or unlawful practices, including those involving AI. This mechanism is particularly relevant where systemic harm arises from the deployment of AI in consumer markets, such as biased credit scoring, discriminatory advertising, or manipulative recommender systems.

Consumer protection and market protection legislation

Many AI systems will be deployed by businesses in interactions with consumers, e.g. to drive sales, improve products or services, engage in customer service or to identify fraud in purchasing transactions. This makes them fully subject to existing consumer protection legislation. Some examples include the Unfair Commercial Practices Directive, the Unfair Contract Terms Directive and the Consumer Rights Directive (CRD). The AI Act declares itself to be complementary to these rules, meaning that even if an AI system is fully brought into compliance with the AI Act it may still violate consumer protection laws.

Most popular GPAI systems are operated by a handful of very large actors, and made available as internet services. Given their size, they are likely subject to the recently-adopted Digital Services Act and Digital Markets Act. The DSA addresses the legal responsibilities of digital services that act as intermediaries in their role of connecting consumers with goods, services, and content. Such intermediaries face limited liability in their connecting role; the AI Act does not change this. Very large online platforms – service providers with over 45 million European users – face thorough transparency provisions, e.g. on content they remove or users sanctioned for inappropriate behaviour. There are also limitations on profiling, targeting children and other platform actions that may be relevant for AI systems.

The DMA aims to ensure fair and open digital markets. It targets large companies that provide core platform services and have a significant impact on the internal market, serve as an important gateway for business users to reach end-users, and have an entrenched and durable position. The tech giants offering AI for business usersmay well qualify as "gatekeepers" under the DMA and face increased scrutiny on their behaviour in the market. One restriction of note is the banning of data sharing between services, meaning an internet platform that offers, say, e-mail and productivity tools cannot use data gathered from user behaviour there to improve an AI system produced as a separate service.

On the topic of (non-personal) data, the Data Act aims to create a fair and competitive data market, ensuring access to and use of data. For AI producers, this means a more equitable landscape for obtaining the data necessary to train sophisticated AI models. It also imposes obligations on data holders and device manufacturers to provide data access to users, which could increase transparency and potentially lead to more innovation in AI services. Data intermediaries and data sharing for the common good is further regulated in the Data Governance Act, on the other hand, focuses on the mechanisms of data sharing and governance. It establishes a framework for data intermediaries and mechanisms for data altruism, where individuals and companies can share data for the common good.

European cybersecurity regulations

Recognizing that cybersecurity is a key aspect of the Digital Decade, the European Union has (or is in the process of) adopting a variety of laws regarding cybersecurity requirements. This follows in the footsteps of – again – the GDPR, which already requires personal data processing to be subject to "adequate technical and organization security requirements". An often-heard critique of this legal requirement is that it is too vague and open-ended. The new regulations seek to provide more certainty through standardization and formal assessments.

Already in 2019, the EU adopted its Cybersecurity Act, providing a scheme for voluntary cybersecurity certification, more on which in chapter 4 (robustness and safety). In 2021, the Cyber Resilience Act (CRA) was proposed. The CRA sets cybersecurity requirements for "smart devices", such as internet-enabled music players, robot vacuum cleaners and so on. Many of these devices contain functionality that meets the definition of AI in the AI Act. If the functionality also qualifies as "high-risk", the AI Act prescribes various security requirements, which can be fulfilled by CRA compliance. This in turn requires a specialized audit by an external party.

December 2022 saw the adoption of the "Directive on measures for a high common level of cybersecurity across the Union", commonly known as the NIS2 Directive. This law sets minimum cybersecurity standards for critical or vital infrastructure in both the public and private sectors. This includes telecommunications, transport and banking, but also food, health and the manufacturing of certain economically important products such as medical devices, computer equipment and heavy or advanced machinery. Involvement of AI in such an infrastructure thus requires a careful evaluation of the NIS2 requirements. The financial sector saw the introduction of the Digital Operations Resilience Act or DORA, seeking to improve its resilience and cybersecurity, although without explicit reference to AI.

From law and ethics to practical assessment

With the AI Act now providing a legal foundation for trustworthy AI in Europe, the central question becomes: how can organizations translate these high-level obligations into concrete practice? To navigate these questions, AI practitioners and compliance professionals need more than law. They need practical tools for ethical self-assessment and governance. Two of the most influential are the *Guidelines for Trustworthy AI* and the ALTAI (Assessment List for Trustworthy AI). These frameworks bridge the gap between legal obligation and ethical aspiration, and provide a roadmap for aligning systems with European values, fundamental rights, and societal expectations.

Lawful, ethical and robust

The *Guidelines for Trustworthy AI* were the direct inspiration for the AI Act. They start with three basic ethical assumptions for what it calls 'trustworthy AI', i.e. AI that is worthy of our trust:

1. AI should be **lawful**, complying with all applicable laws and regulations;
2. AI should be **ethical**, ensuring adherence to ethical principles and values; and
3. AI should be **robust**, both from a technical and social perspective, since, even with good intentions, AI systems can cause unintentional harm.

The lawful nature of AI was not addressed in the Guidelines but is of course now the subject of the AI Act. The Guidelines focused mostly on the ethical aspects of AI, as ethics are usually one step ahead of legislation yet provide some form of guidance on what is desired or acceptable in society. Ethical AI is closely intertwined with the third aspect of robustness: this term does not only refer to a technical perspective (high level of availability, not vulnerable to security breaches, etcetera) but also a social perspective (not having unintended adverse impacts).

Four ethical principles

The Guidelines derive four ethical principles for AI from the fundamental rights that are enshrined in the Charter of Fundamental Rights of the European Union. These closely match the work of AI4People and UNESCO referred to in chapter 1 and are derived from the EU's Charter of Fundamental Rights. The four principles are:

1. **Respect for Human Autonomy**: This principle aligns with Articles 1 and 2 of the Charter, which affirm human dignity and the right to life, respectively. It emphasizes the importance of AI systems supporting individuals' capacity to make their own choices and control their own lives, without undue influence from automated decision-making.
2. **Prevention of Harm**: This principle is reflective of Article 3, which guarantees the right to the integrity of the person, including both mental and physical well-being. It underscores the imperative that AI systems should not cause physical or psychological harm and should be developed with a precautionary approach to risks.
3. **Fairness**: Fairness is a principle that resonates with several articles in the Charter, including Article 20, which ensures equality before the law, and Article 21, which prohibits any discrimination. It calls for AI systems to be equitable and to provide equal treatment and opportunity for all individuals, thereby avoiding biased outcomes.
4. **Explicability**: This principle relates to the rights of transparency and information in governmental functions, as outlined in Article 41, as well as the right to fair processing of personal data (Article 8). Both demand that AI systems be understandable to users, with clear explanations provided for decisions that significantly affect people's lives, ensuring accountability and the possibility of redress.

Seven requirements

For implementors of AI, the Guidelines translate these principles into seven concrete requirements:

1. Human agency and oversight, including fundamental rights, human agency and human oversight.
2. Technical robustness and safety, including resilience to attack and security, fall back plan and general safety, accuracy, reliability and reproducibility.
3. Privacy and data governance, including respect for privacy, quality and integrity of data, and access to data.
4. Transparency, including traceability, explainability and communication.
5. Diversity, non-discrimination and fairness, including the avoidance of unfair bias, accessibility and universal design, and stakeholder participation.
6. Societal and environmental wellbeing, including sustainability and environmental friendliness, social impact, society and democracy.
7. Accountability, including auditability, minimisation and reporting of negative impact, trade-offs and redress.

This list closely follows the fundamental rights from the Charter and matches the ethical points of attention found in the many codes of ethics referred to earlier. While the AI Act does not explicitly refer to the Guidelines, each of the above seven requirements can be found in at least one of its requirements for high-risk AI systems. We will encounter each of these requirements in the later chapters and use them as practical interpretations of the high-level wording from the AI Act's requirements.

Building on the Guidelines is the *Assessment List for Trustworthy AI*, which is meant to guide AI practitioners to achieve Trustworthy AI. It contains a large set of questions designed to prompt stakeholders to include ethical considerations into the design, deployment or use of AI.[12] In this book, the questions from the Assessment List will form an important part of each chapter's tools for applying the AI Act's legal norms in conjunction with the ethical requirements from the Guidelines.

Key takeaways

The AI Act is not an isolated regulatory initiative, but a cornerstone of a broader European strategy. It places risk at the center of its legal structure. Understanding the categories of risk, and the obligations tied to each, is essential for organizations to properly assess, classify, and manage their AI systems. Yet the law alone does not provide all the answers. Tools like the Guidelines for Trustworthy AI and the ALTAI assessment framework offer practical direction on what "robustness," "human oversight," and "ethical design" mean in context. These instruments help transform legal requirements into operational safeguards – bridging the gap between compliance and trustworthy implementation.

In the next chapter, we turn to exactly that challenge: how to operationalize AI compliance inside real organizations. From conformity assessment to documentation, governance roles, and internal procedures, we explore how the legal and ethical expectations introduced in the AI Act can be put into practice – and where human judgment remains irreplaceable.

3

Operationalizing AI Compliance

Now that we have established the legal structure of the AI Act, the focus shifts to its practical application. This chapter addresses the question that every organization working with AI ultimately faces: how do we comply with the Act in real-world settings? While the law provides clear obligations, it leaves many operational details – such as risk classification, role assignment, documentation, and ongoing monitoring – to be worked out internally. This chapter equips you to navigate those grey zones, bridging the gap between legal requirements and organizational practice. Whether you're a provider, deployer, integrator, or compliance officer, the following sections will guide you through the tools, thresholds, and decisions required to make AI systems not only lawful – but governance-ready.

From legal text to practical risk classification

While the primary goal of the AI Act is to manage the risks associated with AI, an implicit secondary goal is to create market certainty. Earlier experiences with technology regulation, notably the GDPR, has shown that open criteria like "likely to result in a high risk" actually hamper compliance and create uncertainty in organisations. The AI Act therefore defined a clear process to determine whether an AI system is high-risk.

Recognizing AI in practice: The first compliance step

Before an organization can assess risk, assign roles, or document obligations, it must answer a far more basic question: are we dealing with an AI system at all? While the AI Act provides a formal definition (see Chapter 2) this definition is broad, and in practice, AI functionality is often hidden, unannounced, or misunderstood.

This is especially true in the context of shadow IT and off-the-shelf SaaS platforms.[1] Many AI systems are not formally developed, procured, or deployed as 'AI.' Instead, they enter the organization through side arrangements, departmental subscriptions, pilot

By the end of this chapter, you'll be able to ...

- Apply the AI Act's classification logic to real-world AI systems.
- Understand the obligations for different operators in the AI value chain.
- Navigate the steps and tools required for compliance.

programs, or embedded features in existing tools. A CRM system might start offering AI-generated lead scores; an HR suite might quietly introduce a CV screening module powered by machine learning; a chatbot plugin may now integrate a foundation model backend. In all of these cases, AI functionality may be active without the awareness of compliance, legal, or IT governance teams.

Worse, suppliers themselves may not clearly signal that a new feature falls under the scope of the AI Act. It is not uncommon for vendors to describe AI-enabled features as "intelligent automation," "smart assistants," or "predictive analytics" – terms that sidestep regulatory language and may prevent buyers from recognizing the legal implications of the product.

This creates a serious compliance risk: if an AI system is present but undocumented, none of the obligations related to classification, oversight, or incident reporting can be fulfilled. For this reason, the first operational step for AI compliance is to establish a discovery and inventory process. Which software or services is used and where – and could it be considered an AI system?

The two-tier discovery process

To ensure full coverage of AI Act obligations, organizations should implement a two-tier process for identifying AI systems across their digital environment. The first tier or phase is a broad discovery or quickscan. It is designed to be inclusive. Its goal is to capture any system that might qualify as AI, without immediately deciding its legal status. False positives thus are to be expected.

This process can cause unrest within the organization.[2] Asking teams whether they are using "unregistered" or "unapproved" AI may come across as accusatory, particularly in environments with low psychological safety or high regulatory anxiety. Some employees may feel the need to defend their use of tools, while others may intentionally withhold information to avoid scrutiny. To mitigate these risks, organizations should frame the discovery effort as collaborative, not disciplinary. Use neutral language such as "automated tools" or "data-driven components" rather than "AI," which may provoke defensiveness.

Questions to ask include:

1. Does the system analyze data and produce outputs beyond static reporting?
2. Is it marketed as "intelligent," "predictive," "adaptive," or "automated"?
3. Does it use natural language, image, speech, or behavioral input?
4. Does it make classifications, suggestions, or decisions?
5. Does the supplier mention machine learning, AI, inference, or models?

Any system that meets one or more of these criteria should be provisionally flagged as "potential AI", and entered into an AI System Inventory for Tier 2 review. This tier is particularly important for identifying shadow AI, such as third-party SaaS tools with unannounced AI features or modules to official systems (such as CRM/HR/ERP) updated with AI functionality.

In the Tier 2 phase, flagged systems are reviewed more closely, either by a compliance officer or legal/technical AI assessor. Assess:

1. Does the system perform inference (i.e. generate outputs based on patterns learned from data)?
2. Was it trained on historical data to predict, classify, or recommend?
3. Does it operate with some autonomy, or affect decisions/outcomes?
4. Is it marketed for high-risk domains (education, employment, law enforcement, health, etc.)?
5. Is it used to profile individuals, as defined under the GDPR?

Corner cases, such as rule-based systems with machine-learned components, or GPAI systems with extensive fine-tuning, should be escalated to the legal team for classification. Only after passing this second tier should a system be deemed in or out of scope.

Once the second-tier analysis is complete, and especially when edge cases have been reviewed by legal or compliance experts, a system can be provisionally deemed in scope, out of scope, or accessory under the AI Act. At that point, the task shifts from identification to documentation. This is where practical compliance begins in earnest: not with regulatory articles, but with a clear understanding of what the system does, how it operates, and what role it plays within the organization. To support that, the next step is to formalize this understanding through the Use Case Card.

The Use Case Card: A compliance entry point

While the AI Act defines high-risk AI systems through formal legal categories (as covered in Chapter 2), compliance begins with something far more practical: understanding and documenting what your AI system actually does. The essential tool for this is the Use Case Card (UCC), a structured format for scoping, documenting, and classifying AI systems under the AI Act.

Introduced by Hupont et al.[3], these cards provide a standardized methodology for documenting AI use cases, focusing on the intended purpose and operational use of an AI system rather than its technical aspects. This mirrors the risk-based regulation of use cases in the AI Act, and in fact the methodology closely matches the various AI Act requirements.

Use case cards differ from traditional model cards[4] in that model cards communicate the specifications, performances, and intended uses of machine learning models as such, while use case cards are designed to frame and contextualize the operational and intended purpose of AI systems. If we consider model cards as the "nutrition label for AI models", providing detailed information about their ingredients (data), nutritional value (performance metrics), and recommended serving size (intended use), then use case cards can be thought of as a recipe book for AI systems. They not only list the ingredients but also describe how the AI system is prepared and served in real-world contexts, including potential variations (different use cases), serving suggestions (intended purposes), and warnings about possible misuse or allergic reactions (foreseeable misuses and risks).

The UCC combines two core elements:

1. A tabular overview of the system's intended purpose, actors, data flows, and legal touchpoints (including AI Act classification);
2. An optional UML diagram, which visualizes system boundaries, users, and interactions.

Used properly, the UCC provides the foundation for:

- Documenting factors determinative of whether the AI system falls under the AI Act's scope;
- Identifying whether it is high-risk, purely accessory, or non-regulated;
- Assigning roles (provider, deployer) across the system's lifecycle;
- Preparing documentation for conformity assessment, technical files, and post-market monitoring.

Who does what? Roles and responsibilities in practice

Once an AI system has been identified and classified, the next compliance task is to determine who is responsible for what. The AI Act defines several legal roles – provider, deployer, importer, distributor, and authorized representative – each with different obligations. But in real-world environments, especially those involving SaaS platforms, GPAI integration, or outsourced development, these roles are rarely straightforward.

Mapping legal roles to real-world functions

The AI Act distinguishes clearly between legal roles, but in real organizations, these categories rarely map one-to-one to job titles or business units. Understanding who is legally responsible under the Act requires translating those legal definitions into operational roles and contractual relationships.

USE CASE		**TalentMatch AI**		
Intended purpose	**Context of use**	The AI application supports the selection process of candidates for open positions by automatically analyzing CVs and cover letters and generating a suitability score. The system is used by HR departments of large employers.		
	Scope	Analyzing application documents and historical personnel data to calculate an aptitude score without human intervention in the first round.		
	SDGs	Decent work and economic growth		
Product (Annex I)		Not applicable	Safety component?	Not applicable
Application (Annex III)		AI systems intended to be used for recruitment or selection of natural persons, notably for advertising vacancies, screening or filtering applications, evaluating candidates in the course of interviews or tests.	Significant harm?	No/Unknown
Primary actor		The HR department of the medium-sized and large company TalentWorks BV		
Stakeholders	Stakeholders	Description		
	HR employees, job applicants, works council, trade union, Labour Inspectorate, Dutch Data Protection Authority.	These groups have an interest in fair, transparent and non-discriminatory recruitment. Applicants in particular should be given equal opportunities, regardless of gender, age or origin.		
Succesful end condition		A timely, fair and well-founded shortlist of candidates for further selection, without systematic bias and with compliance with privacy legislation.		
Failsafe		Interrupting the process in case of missing or unclear data; human assessment. Logging of all analyses and scores. Periodic audits for bias and consistency.		
Activation		The process starts upon receipt of an application via the portal or e-mail. Upload and automatic processing of documents.		
Main course	Step	Action		
	1	Candidate uploads CV and motivation		
	2	AI checks for completeness and reads the documents		
	3	AI calculates aptitude score based on profile and job requirements		
	4	HR receives ranking and explanation		
	5	HR decides who is invited for an interview		
Extensions	Step	Branching action		
	2a	Documents incomplete: candidate receives request for completion		
	3a	AI score unclear: human judgment required		
	5a	Candidate requests access or disputes score (GDPR request)		
Open issues		How transparent is the scoring method for applicants?Are biases (such as age or ethnicity) unintentionally reinforced?How often and by whom are audits carried out on the system?How are GDPR requests for access, correction or deletion handled?		

The primary role is that of the provider, the entity that develops an AI system or has it developed, and places it on the market under its name or trademark. This includes in-house development teams, AI vendors, and increasingly, organizations that fine-tune, rebrand, or repurpose GPAI models for specific domains. Note that "name or trademark" does not necessarily refer to a registered brand name, the purpose of this qualifier is to allow deployers and others to see whose AI they are dealing with.

The deployer is the entity that uses the AI system under its own authority for purposes it sets itself. This role typically falls to business units, departments, or public bodies that integrate an AI system into their daily operations. However, it also applies to an entity that deploys an AI for public use by third parties, such as with an AI service provider that offers subscription-based access to a web app.

Beyond these, the AI Act also defines:

- Importers, who bring AI systems into the EU market from third countries;
- Distributors, who make systems available without altering them;
- And authorized representatives, who assume legal obligations on behalf of non-EU providers.

The AI Act allows for multiple parties to share these roles across the AI system lifecycle. For example, a university research lab might develop an AI model (acting as a provider), a startup might fine-tune and host it (becoming a provider of that system), and a hospital might integrate it into its triage system (acting as a deployer). Each actor has separate, sometimes overlapping, compliance obligations.

To avoid gaps or duplication in legal responsibility, organizations should perform a role-mapping exercise for each AI system. This ensures that all entities involved understand their obligations and that critical tasks like documentation, logging, and post-market monitoring are clearly assigned.

Multi-actor scenarios: Co-provision, outsourcing, and API layers

Many AI systems today are not built from scratch but assembled from components, services, and models provided by third parties. This creates a reality where multiple actors contribute to the development, deployment, and operation of a system, and where the legal responsibilities under the AI Act may be shared, split, or even contested.

One common scenario is outsourced development. A company contracts a software vendor to build a custom AI system, which the company will ultimately use under its brand. Despite the fact that the contractor developed the system, the contracting company is its provider, simply because it puts its name on it and brought it to

market. Conversely, if the vendor continues to host and update the system on its own infrastructure, it may retain the role of provider, while the company becomes a deployer.

Another increasingly prevalent setup involves API-based AI services. Here, an organization integrates an external AI model, such as a language model accessed via API, into its product or internal tools. If the organization uses the API as-is, with limited control over how it functions or what it returns, it is generally considered a deployer. But if the organization builds an entire system around that model, customizes it for a new purpose, or bundles it with additional logic that transforms its output, it may become a co-provider of the resulting system.

This complexity is amplified in the case of general-purpose AI (GPAI). A GPAI model might originate from a large tech firm, be fine-tuned by a consulting firm, embedded into a domain-specific tool by a startup, and deployed in a critical setting by a government agency. Each actor has some level of responsibility under the AI Act – but their specific obligations depend on the extent to which they shaped the system's design, function, and deployment context.

Organizations should therefore evaluate:

1. Who determines the intended purpose of the system as deployed?
2. Who modifies the model or adds decision logic?
3. Who interacts with affected persons, collects user data, or oversees deployment?

Compliance obligations per role

The AI Act assigns specific obligations to each actor in the AI value chain. These roles come with distinct compliance tasks that reflect their level of control over the AI system and their position in the supply chain. The illustration below summarizes the key responsibilities in a typical deployment scenario of a high-risk AI system. (For non-high-risk AI systems, only transparency obligations may exist.)

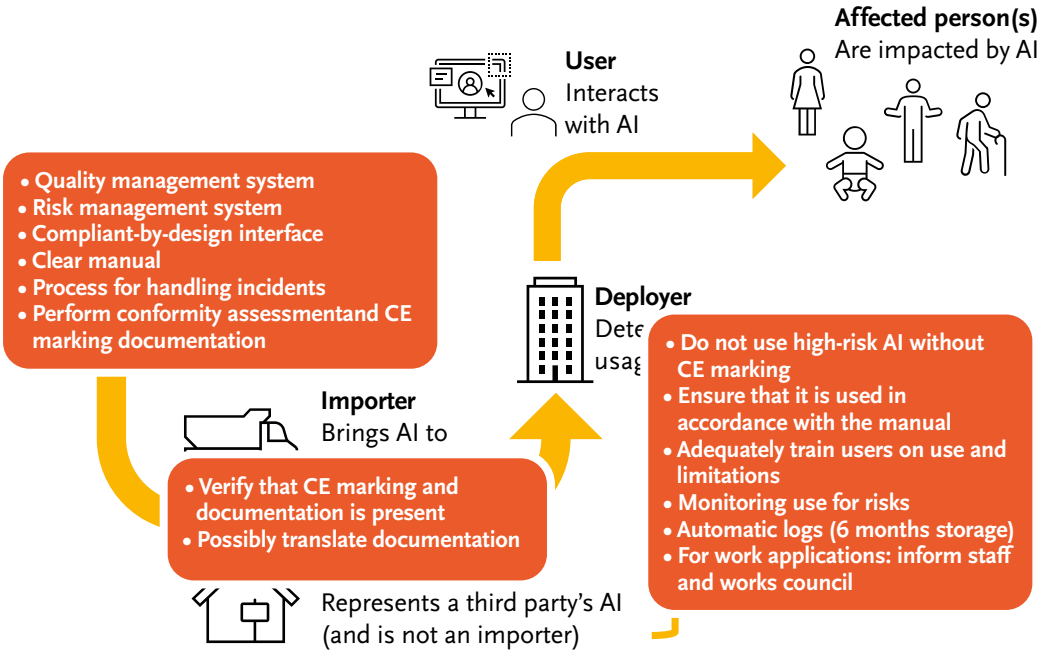
The **provider** is the central figure in AI compliance, bearing the brunt of the Act's compliance requirements and providing key input for the other party's obligations. For high-risk AI, providers are responsible for:

- Establishing a quality management system and a risk management system;
- Ensuring the AI system includes a compliant-by-design interface and a clear user manual;
- Setting up a process for managing incidents and post-market monitoring;
- Performing the conformity assessment before market access (see next section) and affixing the CE marking to indicate legal compliance;
- Enabling transparency and marking of synthetic output (also when the system is non-high-risk).

These obligations form the backbone of AI governance and must be fulfilled before the system can legally be placed on the EU market. Other operators build on the information provided here and may assume that a high-risk AI system is compliant when it bears the CE mark.

The **deployer** is the entity that uses the AI system under its authority, typically as the end user in a business or government context. For high-risk AI systems, deployers must:

- Ensure no AI system is used without valid CE marking;
- Verify that the system is used according to the provider's instructions;
- Train users on system operation and limitations;
- Implement adequate human oversight;
- Monitor the system's use and associated risks;
- Keep automatic logs for at least six months;
- Inform workers and works councils (OR) when AI systems are deployed in the workplace, especially those that may affect conditions or decision-making;
- Ensure the AI system operates with adequate transparency, and synthetic output is marked as such (this also applies when the system is non-high-risk).



Importers bring AI systems from outside the EU onto the EU market. Their role is to act as a gatekeeper, verifying that:

- The AI system has a valid CE mark;
- The required technical documentation is complete and available (importers may provide translations into local languages.

An **authorized representative** acts on behalf of a non-EU provider and assumes legal responsibility for the AI system in the EU. This role is not the same as that of an importer, although one entity may serve as both. Like the importer, the representative must verify the CE marking of the high-risk AI system, but also serves as contact point for market surveillance bodies.

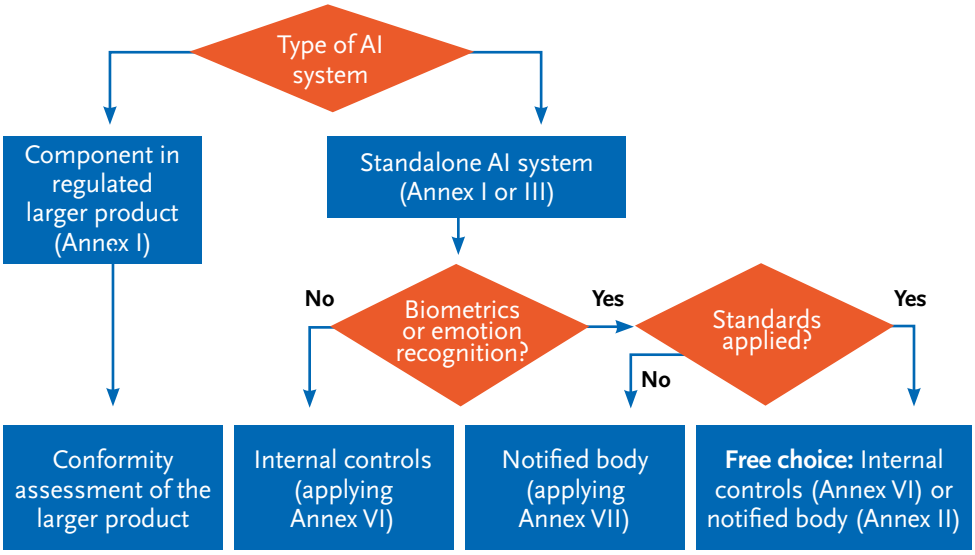
A **distributor** is any actor in the supply chain other than the provider or importer who makes the AI system available. Distributors must ensure that the CE marking and documentation are present and may translate documentation as needed for the target market of deployers. Distributors must be prepared to take corrective action if they learn that a system is non-compliant.

Conformity assessment and technical documentation

Conformity assessment (CA) is the formal process by which a provider demonstrates that an AI system meets all the mandatory requirements of the AI Act before it is placed on the market or put into service.[5] It is a regulatory obligation, not a discretionary risk analysis, and differs fundamentally from assessments like DPIAs (Data Protection Impact Assessments) or FRIAs (Fundamental Rights Impact Assessments), which aim to evaluate and mitigate contextual or rights-based risks. In contrast, CA is objective, procedural, and product-focused: it results in a declaration of conformity and, where required, the application of a CE marking.

Establishing conformity and the role of standards

Conformity assessment under the AI Act is meant to be objective and reproducible. To make that possible, the Act relies heavily on the concept of harmonized standards: technical norms developed by the EU standardization CEN and CENELEC. These carry a so-called presumption of conformity – if a provider follows the applicable harmonized standards, the system is presumed to meet the corresponding legal requirements.



At the time of writing, no harmonized standards for the AI Act have been formally adopted. Dozens of draft standards are under development, ranging from risk management, data quality and transparency to post-market monitoring, but they are not yet binding or available for formal use in conformity assessments.

Voluntary standards and best practices, such as ISO/IEC 42001 (AI management systems) and ISO/IEC 23894 (AI risk management) are available,[6] but as these are not harmonised standards their compliance does not confer a presumption of conformity. Further, the AI Office coordinating the EU's AI efforts has repeatedly signaled that ISO 42001 is not in line with various aspects of the AI Act.

Annex VI of the AI Act, which provides a fallback option: it outlines how to conduct a conformity assessment directly against the AI Act's essential requirements, without relying on external standards. This is a more demanding path, as the burden is entirely on the provider to demonstrate that its system meets the law. There is no presumption of conformity either. Yet, a high-risk AI system cannot be put on the market without a conformity assessment having been performed.

Annex VI requires three items to be verified:

1. The established quality management system is in compliance with the Act's requirements (article 17).
2. The technical documentation covers all aspects for high-risk AI systems (Chapter III, Section 2).
3. The design and development process of the AI system and its post-market monitoring is consistent with the technical documentation.

Conformity assessment routes

There are two primary routes for conformity assessment under the AI Act, depending on the type and risk level of the AI system: the Internal Control Route and the Notified Body Route. For most high-risk AI systems, the provider may follow the internal control procedure. This means the organization conducts a self-assessment of its system's compliance with all applicable AI Act requirements, ranging from risk management and data governance to human oversight and transparency. The provider must:

- Prepare a technical documentation file (see next subsection);
- Conduct an internal conformity assessment against either Annex VI or a harmonised standard (see previous subsection);
- Issue a declaration of conformity;
- Affix the CE marking before placing the system on the market.

In certain cases, however, the AI Act requires an external assessment by a notified body. This is a third-party, government-accredited organization tasked with evaluating compliance. Currently, this option is prescribed for AI systems applying the high-risk use cases of remote biometric identification, biometric categorisation and emotion recognition when no harmonised standard is applied. For other types of high-risk AI the option is possible but not mandatory.

The notified body will examine the technical documentation, risk management procedures, testing outcomes, and human oversight mechanisms. If satisfied, the body issues a certificate of conformity, enabling the provider to proceed with CE marking.

Building the technical documentation file

Part of the conformity assessment process is the requirement to prepare and maintain a technical documentation file. This file demonstrates how the AI system complies with the requirements of the AI Act and serves as the primary reference point for supervisory authorities or notified bodies reviewing the system. Without it, CE marking cannot be lawfully affixed, and the system cannot be placed on the market or put into service.

The documentation must be comprehensive, coherent, and up to date, and must reflect both the design of the AI system and the processes used to ensure its ongoing compliance. The structure and content of the technical documentation are defined in Annex IV of the AI Act. In summary:

- A clear description of the system and its intended purpose, including its use environment and any limitations;
- Performance metrics and a description of the system's behavior under expected conditions;

- Data governance details, including dataset sources, representativeness, and known limitations;
- Design and development procedures, including versions, updates, and testing protocols;
- Documentation on the risk management process;
- Human oversight measures for prospective deployers;
- Technical solutions for transparency, traceability, and explainability;
- Information on post-market monitoring, including logging, feedback loops, and update procedures.

The file must be actively maintained throughout the lifecycle of the AI system, including after deployment. If the system is retrained, fine-tuned, significantly updated, or repurposed for a new intended use, the documentation must be revised accordingly. Failure to do so can invalidate the conformity declaration.

In many organizations, the technical documentation will draw from existing quality management systems, security frameworks, or software development lifecycle records. However, it must be assembled and formatted in a way that explicitly maps to the AI Act's requirements, either through harmonized standards (once available) or direct compliance via Annex VI.

Assessing fundamental rights: The FRIA tool

Impact assessments are systematic evaluations designed to understand the potential consequences of a project, policy, or system. Risk impact assessment models are often used as a tool to consider and preserve data subjects' fundamental rights, as well as ethical and social standing.[7] The AI Act builds on the GDPR's earlier data protection impact assessment by introducing the *fundamental rights impact assessment* or FRIA instrument.

Purpose of the FRIA instrument

The purpose of a FRIA is to "efficiently ensure that fundamental rights are protected", as the AI Act puts it. It should identify the processes in which the high-risk AI system will be used in line with its intended purpose, list time and frequency of use, and examine the people (or groups) that are likely to be affected by it, as well as the specific harms that may occur. This is wrapped up by identifying human oversight measures and specific steps to be taken when the risks materialise, such as complaint mechanisms.

The FRIA was added to the AI Act in response to concerns that the formal process of conformity assessment (see previous section) would be inadequate to address the often-intangible risks to fundamental rights, such as free speech, non-discrimination,

protection of personal data and privacy, and fair treatment by government. The FRIA is open-ended in that it asks to consider *all* fundamental rights, which even includes impact on the environment (compare chapter 9).

The FRIA is not a one-off form to tick off before deployment, but rather is to act as a living rights analysis. Once the AI system is in use, post-market monitoring (see next section) becomes the key instrument to detect changes in risk, unforeseen impacts, or shifts in deployment context. These findings may directly affect the validity of the original FRIA. Revising the RIA thus ensures fundamental rights are continuously protected throughout the system lifecycle.

When is a FRIA required?

The FRIA obligation exists prior to deployment, as deployers are in the best position to evaluate risks to fundamental rights when they take an AI system into use. But to avoid administrative overhead, not all deployers of high-risk AI are obliged to perform a FRIA. First, the requirement applies only when the AI system is high-risk because it concerns a use case listed in Annex III – high-risk due to being a regulated product is outside scope (see chapter 2 for the high-risk classification process). Second, the deployer must be one of a specific category:

- **Bodies governed by public law.** This generally refers to central and local government agencies, but can also cover entities such as national sports federations.
- **Private operators providing public services.** This term is more general than "bodies governed by public law". A public service is an economic activity of general interest defined, created and controlled by the public authorities. In EU policy jargon these are referred to as Services of General Interest (SGI), Examples are education, healthcare, social services, housing and administration of justice. Public services differ from activities by bodies governed by public law in that the latter do not provide economic activities. A court issues judgments; a legal clinic provides the service of legal advice.
- **Creditworthiness checks.** Entities performing creditworthiness checks (but excluding fraud detection) must perform a FRIA prior to deploying that process. This includes agencies calculating credit ratings for others that make the actual decision.
- **Risk assessment and pricing for life and health insurance.** Similar to credit checks, a FRIA is necessary prior to deploying risk assessment and pricing AI systems for the two high-risk insurance use cases of life and health insurance.
- **Law enforcement authorities intending to use a real-time remote biometric identification system in publicly accessible spaces.** This requirement was added as part of the safeguards for the controversial use of real-time biometrics in law enforcement.

Other entities are of course free to do a FRIA, but are not legally required to do so. Note that government agencies must do a FRIA prior to *any* high-risk AI deployment (except cybersecurity, as this is covered in other legislation), while the other entities mentioned must do so only prior to deploying the listed high-risk AI.

Structure of the FRIA

The AI Act deliberately leaves the structure and format of a FRIA open, although the AI Office is tasked with drawing up a template.[8] This gives organizations flexibility, but also creates uncertainty: how do you assess something as broad and abstract as "fundamental rights risk"? Fortunately, several existing instruments offer practical starting points. While none of them was designed specifically for the AI Act, each brings valuable perspectives and components that can be adapted for use in a FRIA.

These tools can be used individually or in combination, depending on your organizational maturity, domain, and risk profile. For instance, public-sector deployers might combine ALTAI for ethical framing with AIIA to structure procedural obligations.

Instrument	Origin & Purpose	Strengths for FRIA	Limitations
ALTAI (Assessment List for Trustworthy AI)	Developed by the EU High-Level Expert Group as an ethical self-assessment tool.	Aligns closely with EU values and principles; strong on fairness, accountability, human agency. Good for identifying categories of risk and ethical gaps.	Not legally framed; lacks concrete criteria for assessing severity or legal justifiability. Needs adaptation for compliance purposes. Very large.
AIA (Algorithmic Impact Assessment – Ada Lovelace Institute)	UK-based framework emphasizing transparency, public engagement, and procedural justice.	Excellent for structuring stakeholder involvement and surfacing social risks. Strong on contestability and governance context.	Lacks operational detail. Not tailored for EU law. May be too abstract or deliberative for organizations seeking a checklist.
AIIA (AI Impact Assessment – Dutch Government, version 2)	Dutch ministry of Infrastructure and Water to assist deployment of trustworthy AI.	Explicit focus on EU fundamental rights. Two-part structure on ethics and technical implications. Covers lifecycle from idea to deployment. Strongly aligned with AI Act.	Designed for public sector ussage. Significant focus on bias and fairness, other risks are treated lightly.
FRAIA (Fundamental Rights and Algorithm Impact Assessment)	Dutch ministry of the Interior and Kingdom Relations to facilitate interdisciplinary dialogue on algorithmic systems.	Facilitates direct use, focuses on process & actors.	Designed for public sector concerns. Over 99 pages. No specific focus on AI risks, no explicit AI Act alignment.

FRIA versus DPIA

The FRIA is quite similar in purpose to the GDPR's Data Protection Impact Assessment (DPIA). Both are mandatory, risk-based assessments under EU law. Both aim to prevent harm to individuals before a system is deployed. And both are required for many high-risk AI applications. But their legal basis, scope, and focus differ significantly.

Aspect	FRIA	DPIA
Legal Basis	AI Act (Article 29, Annex III)	GDPR (Article 35)
Trigger	High-risk AI system by public body or listed private actor	Any data processing likely to result in a high risk to individuals' rights and freedoms
Focus	All fundamental rights, including equality, due process, autonomy, freedom of expression, and protection of environment	Protection of personal data, privacy rights, and other rights and freedoms of persons
Format	Open-ended analysis, template by AI Office to be drawn up	Open-ended analysis, template by EDPB available
Outputs	Risk identification, mitigation plan, human oversight structure, complaint mechanism, governance arrangements	Risk analysis, mitigation measures, may prompt prior consultation with supervisor
Repetition	Must be updated if PMM reveals new risks	Must be updated if processing changes or new risks emerge

Although the trigger requirements are different, in high-risk use cases involving personal data (which includes most of Annex III), both a FRIA and a DPIA may easily required. In such a case, the AI Act prescribes that a single instrument can be used to address both requirements.

Ongoing compliance and post-market monitoring

Compliance with the AI Act does not end at the moment of CE marking or market entry. For high-risk systems, the provider remains responsible for monitoring how the system performs in real-world conditions, detecting any adverse outcomes, and responding appropriately. These duties are known as post-market monitoring obligations and require structured processes for incident logging, change control, risk tracking, and reporting to supervisory authorities.

Post-Market Monitoring

The phrase "post-market monitoring" or PMM refers to all activities carried out by providers to collect and review experience gained from the use of AI. The purpose is to be able to identify any need to immediately apply any necessary corrective or preventive actions. All providers of high-risk AI systems must establish and maintain a post-market monitoring system, a structured way to perform these activities. Size and manner of execution are left to the provider and can differ based on the nature of the AI technologies and the risks of the high-risk AI system. The concept is borrowed from the post-market surveillance for medical devices.[9]

PMMs are the complement to the conformity assessment process. CAs are performed in advance and evaluate the system in the general case. PMMs are performed while the system is in use and evaluates what's actually happening. This enables providers to take immediate action when any serious incident or malfunctioning that constitute a breach of Union law occurs (see next section), or to identify potential deviations from the conformity assessment. Providers are obliged to take immediately any corrective actions needed to bring the AI system under conformity or withdraw it from the market.[10]

At minimum, a post-market monitoring system should include:

1. Ongoing performance evaluation, including monitoring of key outputs and error rates;
2. Risk re-assessment procedures, to re-evaluate known risks and detect emerging ones;
3. Incident detection and escalation mechanisms, including thresholds for internal alerts and regulatory reporting;
4. Data collection mechanisms, such as user feedback loops, audit logs, and system usage analytics;
5. Update tracking, especially in cases where the AI system is retrained, reconfigured, or adapted after deployment.

This process must be documented and integrated into the provider's overall quality management system. The monitoring data should also feed into decisions about whether the system needs to be adjusted, re-assessed, or even re-certified. Deployers may not be responsible for maintaining the PMM system themselves, but they often serve as its primary source of input. This includes reporting anomalies, escalating complaints, and maintaining local logs or performance dashboards. For this reason, clear coordination protocols between providers and deployers are essential and should be agreed upon in the contracts between them.

Serious incident reporting

Deserving particular attention in a PMM system is the identification and reporting of what the Act calls "serious incidents" involving high-risk AI systems. A serious incident is defined as any malfunction, performance issue, or unintended outcome that (directly or indirectly) causes:

1. Death or serious harm to the health or safety of a person;
2. A serious and irreversible disruption of the management or operation of critical infrastructure;
3. A severe or large-scale infringement of obligations under Union law intended to protect fundamental rights;
4. Serious harm to property or the environment.

Providers must report such incidents to the relevant market surveillance authority without undue delay and no later than 15 days after becoming aware of it. After this initial report, the providers must perform the necessary investigations in relation to the serious incident and the AI system concerned. This includes performing a specific risk assessment of the incident, and corrective action such as a recall of affected products or a change to the AI system.

For deployers, the obligation is more indirect: if the deployer becomes aware of a serious incident, they must promptly inform the provider. If this does not have the desired effect, they must contact the market surveillance authority as if they were the provider. Moreover, deployers must monitor the use of high-risk AI systems for potential risks and proactively inform the market surveillance authorities.

To comply with this obligation, providers should establish:

1. A dedicated incident reporting channel that reaches the responsible compliance officer;
2. Clear criteria for what qualifies as a reportable incident, ideally aligned with internal risk registers;
3. A structured response process for investigating the root cause, assessing legal impact, and documenting remediation;
4. Training and awareness for both technical and non-technical users to recognize potential incidents and trigger escalation.

Logging and traceability mechanisms

Logging is one of the most underappreciated compliance mechanisms under the AI Act. Providers of high-risk AI systems must facilitate automatic logging of events during system operation, to ensure that the system's functioning can be traced, reviewed, and, if necessary, audited. Deployers in turn must activate such logging and retain them for six months (shorter if GDPR requirements would otherwise be violated).

Logging serves two key purposes:

1. It supports accountability and explainability, enabling human reviewers, auditors, or regulators to reconstruct how a given output was generated.
2. It underpins post-market monitoring by making it possible to detect patterns of failure, misuse, or degradation over time.

To comply with the AI Act, logging must be:

1. Automatic: The system must generate and store logs without relying on user intervention;
2. Comprehensive: Logs should cover the sequence of inputs, relevant decision-making steps, and output actions;
3. Secure and tamper-proof: Logs must be protected against unauthorized access, modification, or deletion;
4. Retained for a minimum of six months, unless otherwise required by sectoral regulation.

The AI Act does not prescribe a specific technical format, but the logs must be structured in a way that supports traceability.[11] That means they should link inputs to outputs, include timestamps, system states, and model version identifiers. For systems using GPAI components, logs should also document external dependencies – such as API calls or embedded model versions. Traceability also requires internal version control across system updates, model retraining, or interface changes. Each version of the AI system that is deployed should be uniquely identifiable, and the technical documentation should indicate which version was live at any given time.

For deployers, logging is just as important. Maintaining usage logs can help detect misuse, confirm that outputs were acted upon appropriately, and support internal investigations or user complaints. Where AI systems are deployed in decision-support contexts, the logs can also show whether human reviewers exercised meaningful oversight.

Supervision and enforcement

Enforcement of the rules of the AI is grounded in the EU's long-standing approach to product regulation.[12] Rather than establishing a single, centralized AI regulator, the Act establishes a network of sectoral *market surveillance* authorities in each Member State, with the central AI Board providing coordination and exchange of information. In addition, given the importance of fundamental rights being protected in the era of AI, national oversight bodies for these rights are given new powers. For GPAI the European Commission itself is the competent supervisor.

Market surveillance as the basis

The primary means of enforcement under the AI Act is built on the market surveillance system already in place for regulated products across the EU.[13] Each Member State designates one or more national Market Surveillance Authorities (MSAs) to oversee compliance with the Act within their jurisdiction. These authorities are responsible for monitoring high-risk AI systems that are placed on the market or put into service, carrying out inspections and audits, investigating complaints, and imposing corrective measures where necessary.

While each member state makes their own choices, the general approach is that market surveillance of AI is performed by the already-existing market surveillance framework for regulated products such as medical devices, machinery, foodstuffs and consumer products. MSAs are expected to leverage their existing structures and expertise in technical conformity, documentation review, and enforcement actions. Where no specific existing authority is available, typically the GDPR's data protection authority assumes the role of MSA.

MSAs have the power to:

- Conduct on-site and remote inspections;
- Request technical documentation and access to logs;
- Under strict confidentiality, gain access to training data and model parameters;
- Order the suspension or withdrawal of non-compliant systems;
- Impose administrative fines or other corrective measures;
- Coordinate enforcement in multi-jurisdictional deployments.

MSAs are expected to draw up guidelines for enforcement and the manner in which fines are assigned. The AI Act provides a level of guidance for the height of fines:

- Applying a prohibited practice is the highest category of fine: up to EUR 35 million, or 7% of worldwide annual turnover if higher.
- Failure to adhere to an operator's basic provisions such as risk management or post-market monitoring, or the transparency requirements (see chapter 7): up to 15 million, or 3% of worldwide annual turnover if higher.
- Supplying incorrect, incomplete or misleading information to MSAs or notified bodies: up to 7,5 million, or 1% of worldwide annual turnover if higher.

The needs and positions of SMEs must be considered critically when imposing fines. To this end, the AI Act in fact uses the two numeric criteria in reverse: for SMEs the maximum fine is the *lower* of the monetary amount and the percentage of annual turnover. Issuance of fines is of course in all cases subject to the member states' procedures of administrative law and can be challenged in court.

Protecting rights beyond compliance

As noted in chapter 1, many AI systems engage fundamental rights protected under EU law. To ensure these rights are enforceable in practice, the AI Act preserves and reinforces the role of national authorities and bodies tasked with protecting fundamental rights, alongside the technical oversight performed by the MSAs.[14]

These bodies, such as equality commissions, ombuds institutions, and data protection authorities (DPAs), retain full autonomy to investigate whether an AI system infringes rights such as non-discrimination, data protection, access to education, or fair treatment in employment. The AI Act even gives them further powers to request and access any documentation when necessary to effectively fulfil their mandates within the limits of their jurisdiction. For instance, a national equality body could investigate whether an AI-based hiring platform disproportionately filters out candidates with disabilities or minority backgrounds, and request access to the system's data quality assessments, fairness testing results, and logs of past decisions to assess whether indirect discrimination has occurred.

Fundamental rights authorities or FRAs do not have direct enforcement or punitive powers like MSAs have. With their investigate powers, they can acquire what they need to write a report or finding, but cannot do more than publish it or urge the organization to change their ways. However, they can of course present their reports to the competent MSA which does have the power to issue a fine.

Examples of rights-focused enforcement bodies include:

- The French Defender of Rights (Défenseur des droits), overseeing discrimination, children's rights, and digital fairness;
- The Dutch Institute for Human Rights, which in 2022 investigated AI systems for student proctoring in the wake of COVID-19;
- The Irish Ombudsman for Children, focused on rights of minors.

A special mention goes to data protection authorities. Of course these agencies have supervisory powers under the GDPR. But as the GDPR is directly related to the fundamental right of protection of personal data, the AI Act recognizes them also as an FRA. This gives them additional powers of investigation when a GDPR violation is hard to establish.

The Commission and GPAI oversight

Due to the importance of GPAI models in today's AI systems, and the fact that most GPAI is provided by a handful of (American) companies, the drafters of the AI Act have chosen to have supervision of GPAI handled directly at the EU level. The European Commission has established the AI Office as its executive body for coordinating and

exercising supervision over GPAI models, particularly those presenting systemic risks. The AI Office is empowered to monitor and enforce compliance by GPAI model providers. To this end, it has the power to request documentation and more generally any information it deems necessary – with a separately-established panel of scientific experts to help it make sense. Next, the Office can conduct evaluations of any GPAI models on the market and request cooperation and access to source code to this end.

When an AI Act violation or systemic risk is found, the Commission takes over. Its powers include the ability to order interim or corrective measures, including restrictions on use or changes in model design. The Commission may impose fines up to 3% of their providers' annual total worldwide turnover, or EUR 15 million if higher, if it finds a violation of the AI Act or a failure to cooperate or provide information.

When a provider develops both the GPAI model and the AI system built on top of it, the AI Office assumes MSA powers over the system as well. This ensures that model developers cannot avoid scrutiny simply by reclassifying their outputs as downstream tools. In such cases, the AI Office has all the powers of a national market surveillance authority, including access rights, enforcement tools, and investigatory powers.

AI literacy as a compliance obligation

Compliance (and governance, see chapter 11) is not only technical and legal: it is also cognitive and cultural. An AI system cannot be operated safely if the people involved in its development, deployment, or oversight do not understand what it is doing.[15] Hence the AI Act's requirement for AI literacy.

Defining AI literacy and who it applies to

AI literacy should equip providers, deployers and affected persons with the necessary skills, knowledge and understanding to make informed decisions regarding AI systems, as well as to gain awareness about the opportunities and risks of AI and possible harm it can cause. This is justified by the fast rise and enormous impact that AI can have on society and the people in it. The AI Act thus obligates both providers and deployers to ensure a "sufficient level" of AI literacy among staff and other persons interacting with AI systems on their behalf.

The Act defines AI literacy in functional terms: a combination of skills, knowledge, and understanding that allows individuals to use, interpret, monitor, and challenge AI systems responsibly. The goal is to empower actors throughout the AI value chain to ensure appropriate compliance and enforcement. This includes technical staff, operational users, management, and oversight roles.

The obligation explicitly applies to:

- Providers, who must ensure that development and testing teams understand the AI system's technical capabilities and legal constraints, and who must ensure their AI systems (both high-risk and not) come with adequate instructions and fail-safes against accidental misuse;
- Deployers, who must ensure that operational users, risk managers, and internal overseers are capable of interpreting the system's output and maintaining appropriate oversight.

The literacy requirement is contextual and relative. It must be tailored to the user's role, background, and training, the type of AI system and the people affected by the system's outputs, including vulnerable or marginalized users.

Some organisations, fearful of the risks that AI poses, issue a company-wide ban on AI use. However, that's not the same as AI literacy. Literacy requires positive engagement, and thus training, awareness-building, internal policy guidance, and the capacity to identify and escalate issues. Blanket prohibitions may reduce surface risk but create shadow IT practices, suppress organizational learning, and leave teams unprepared for inevitable exposure to AI technologies.

Building and demonstrating AI literacy

The AI Act does not prescribe how to achieve literacy, nor does it set formal penalties for failing to do so. However, it requires that human overseers (see next chapter) be appropriately trained. What's more, the article on AI literacy is formulated as a legal right of employees – allowing them to demand literacy training. Customers may demand "AI Act compliance" in purchasing agreements. All this means AI literacy is unavoidable. But how to implement it?

Some Member States have taken initiative. The Dutch Data Protection Authority, for example, released a four-step model:

1. identify AI systems and users,
2. set learning goals,
3. implement training strategies, and
4. evaluate progress over time.

This approach mirrors traditional organizational change frameworks and offers a scalable approach for larger institutions. It does enable adaptation to new developments and can be adjusted to the needs of specific groups within the organisation.

Key building blocks for compliance include:[16]

1. Inventories of affected roles and AI system touchpoints;
2. Role-specific training modules, especially for users of high-risk AI;
3. Onboarding protocols that incorporate AI ethics, bias awareness, and reporting duties;
4. Documentation of training activities, to support audits and incident response reviews.

At the EU level, the AI Office maintains a living repository of AI literacy practices, accessible via its website as of February 2025. This repository includes use cases, sectoral toolkits, and sample training materials drawn from industry, public administration, and education.

Key takeaways

This chapter translated the AI Act's legal obligations into operational practice: from identifying AI systems and assigning legal roles, to building documentation, conducting conformity assessments, and maintaining post-market controls. We saw how enforcement is distributed across national authorities and the EU, and how AI literacy is a formal, ongoing responsibility. With this compliance structure in place, we now turn to the first of the Act's core ethical dimensions: ensuring that AI systems support human agency and oversight.

4

Reinforcing Human Agency and Oversight in AI

Amidst the rise of autonomous systems, it is essential to understand the continuing role of human involvement in AI. From design to deployment, the concept of human agency, meaning the ability of individuals to make independent and informed choices, plays a central role in shaping the societal impact of these technologies. The AI Act to this end requires high-risk AI systems to include effective forms of human oversight, both to prevent harmful over-reliance and to ensure that humans remain meaningfully involved in decisions with serious consequences. This chapter explores how interaction between people and AI systems can influence autonomy, trust, and accountability, and offers practical strategies for integrating oversight in both design and operation.

Understanding Human-AI Interaction

Today, AI is playing a pivotal role in a wide array of applications, from trivial tasks such as recommending a movie to significant decision-making processes in fields like healthcare, finance, and transportation. As noted in chapter 1, a key aspect of AI is its autonomy: the ability to produce interactions, classifications, decisions and the like without human involvement. This rise in autonomous computing technologies has a direct impact on human autonomy.[1] In this section, we will discuss the design of AI systems that are meant to interact with, guide, or make decisions for human end-users. The focus will be to understand both the positive and negative impacts and implications of such interaction.

The rise of computer interaction

The freedom and autonomy of its citizens should be a key priority for each democratic society. Automation and information technology has clearly contributed to this priority: when menial tasks can be carried out by machines, humans can spend their energy on more high-level tasks. An oft-cited example is that of the washing machine and vacuum cleaner, which significantly increased women's opportunities to engage in work and express themselves by freeing up time previously consumed by household chores.[2]

By the end of this chapter, you'll be able to ...

- Formulate the key role of human agency in the deployment of AI.
- Mitigate potential risks to human autonomy and oversight.
- Suggest effective human oversight and response mechanisms in AI systems.

ALTAI

A1. Is the AI system designed to interact, guide or take decisions by human end-users that affect humans or society?

AI comes with the promise of doing the same but for cognitive tasks: driving cars, executing routine tasks, suggesting new music or food, and so on. However, a key difference with the previous example is that such systems are more involved in decision making, rather than simply carrying out the wishes of their human operators. A classic saying in IT is "computers do what you say, not what you want". In AI, this can be reformulated as "computers do what they presume you want".

That is not to say all AI, or even all automation is necessarily threatening to society's wellbeing. The initial reason for the introduction of automation was to reduce the workload of operators, and thus reduce operational costs and errors, while increasing accuracy.[3] Additionally, automation and computer-implemented safeguards have significantly increased safety in many domains, such as transportation. This shift from performing tasks to supervising automation introduces a tension: while it may increase efficiency, it can also reduce the individual's sense of control and influence. Without safeguards, this undermines human agency, which is now recognized as a key consideration in the regulation of high-risk AI systems.

A key insight however is that the rise in automation has delegated the nature of human activity from direct manual control to partial or total supervision.[4] The negative impact of such a change on operators' activities is referred to as the phenomenon of *out-of-the-loop performance problem* (OOTL). Decreased vigilance, complacency or overconfidence in the system's capabilities, and a loss of situational awareness on the part of the operator have been identified as factors that may contribute to this phenomenon.[5] These effects are now widely acknowledged in AI governance frameworks, which require that systems be designed to maintain human involvement and uphold users' capacity to make independent judgments.

The importance of agency

The concept of 'agency' has often been identified as a potential bridge to improved human performance. Essentially, the "sense of agency" is our conscious perception of initiating and managing our own actions, crucial for everything from motor control to social interactions and a driving force behind human behavior. In the field of psychology, various factors have been suggested as contributing to the impact of AI on agency:[6]

1. Decrease in the sense of agency when interacting with highly automated systems is likely to seriously threaten the acceptability of the system's decisions by human operators.

❷ Any change in self-agency can modulate the operator's involvement in the task at hand.
❸ A change in the feeling of agency could have a direct influence on cognition, and through this, on operational performance.

Studies suggest that it is not the level of automation itself that modulates the sense of control in the first place, but rather the amount of control remaining over the action.[7] An AI assistant that rewrites a short prompt into a flowery e-mail has full autonomy over what to write, but leaves all control to the human – he or she can rewrite as desired prior to sending. An AI app that tells a human food delivery driver where to go and when to be there, on pain of not getting paid for the delivery, reduces almost all control and thus has a much more negative impact on the sense of autonomy.

ALTAI

A1a. Could the AI system generate confusion for some or all end-users or subjects on whether a decision, content, advice or outcome is the result of an algorithmic decision?

The ultimate reduction of control is the oft-cited "social credit system" of the People's Republic of China. In this system, human activity is constantly monitored and those who are found to violate certain rules or norms are awarded demerits (or credits, for good behavior). A sufficiently low level of "social credit" impacts legal rights – one may no longer travel by plane, for instance.[8] This system requires constant mass surveillance, including biometric identification to keep track of all infractions.

Social credit systems exhibit a key factor to loss of agency, namely that of unpredictability – factors that impact one's credit score may change from week to week, as well as the amount of points that the impact has. What's more, an offender is not addressed by a police officer or other official, nor does he or she get a chance to dispute the demerits awarded. This tremendous impact on human agency, dignity and fundamental rights is the reason the AI Act explicitly calls out social credit systems as prohibited practices, as well as real-time biometric surveillance at a distance (on which subject more in chapter 6).

The prohibition of social credit scoring is actually not the first practice in the AI Act – that honor goes to subliminal techniques for manipulating human vulnerabilities or unconscious thought. This is an extreme example of an AI system overriding, "distorting" in the AI Act's terminology, of an operator's sense of control or agency towards the system. Subliminal messaging has been the subject of controversy since 1957, when researcher James Vicary hid the message "Hungry? Eat popcorn!" in a movie shown in a New Jersey movie theatre and claimed it increased popcorn sales by 57%.

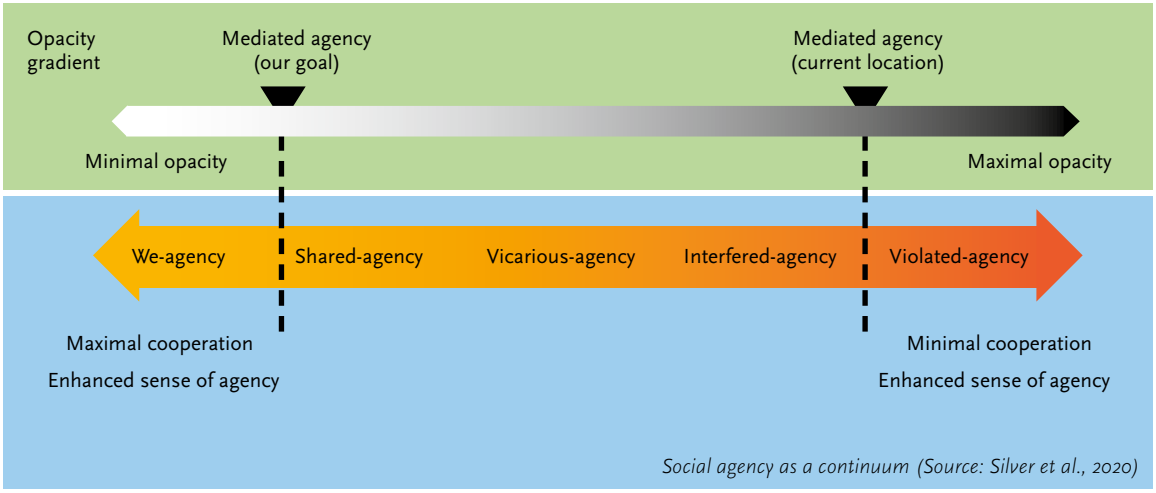
Subliminal text messages embedded in images created with the Midjourney generative AI system. Can you spot the messages?

While Vicary's results were never duplicated, and no other study could ever show significant effect of subliminal advertising,[9] the concerns over manipulation of the unconscious by hidden messaging has remained on legislators' agendas ever since. The prohibition has been criticized as superfluous: the scope overlaps heavily with other legislation such as the Audiovisual Media Services Directive (2010), the Unfair Commercial Practices Directive (2015) and the Digital Services Act (2022) and no AI system currently on the market is even alleged to employ subliminal messaging.[10]

Agency and cooperation

Agency can be thought of not only as an individual state, but also as a relational one. The quality of cooperation between a human and an AI system can shape how much agency the user retains or feels. The image on the opposite page illustrates various forms of agency as they relate to transparency (opacity).[11] On the left side, we see the maximal form of cooperation and agency: we-agency, where co-actors share a common identity and a common goal, with a strong feeling of cooperation ("as a team"). A lesser form is shared-agency, where the agents work together but are clear about their separate roles. Vicarious agency occurs when the result of another agent's action is wrongly attributed to the self, and interfered agency occurs when goals are ill-defined, or when there is no cooperation, or when the actions of the other agent are unpredictable. In this case, the presence of another agent interferes with our own agency. Finally, in "violated-agency" an individual feels a loss of control over their actions, often perceiving their actions as being influenced or manipulated by an external entity.

A clear example of violated agency can be seen in app-based gig work, where drivers receive assignments and navigation routes from an opaque algorithm. The system may penalize deviations or fail to explain why some jobs are assigned and others withheld.



Social agency as a continuum (Source: Silver et al., 2020)

Over time, this can erode the worker's ability to make autonomous decisions, especially when refusal leads to lost income or account deactivation. The individual remains formally in control but is functionally constrained by a system they cannot question or influence.

ALTAI

A1b. Are end-users or other subjects adequately made aware that a decision, content, advice or outcome is the result of an algorithmic decision?

The way an AI behaves thus is crucial to an adequate sense of agency in the human operator. This is where the often-cited complaint of "black box" behavior comes from: when the AI merely delivers a conclusion without justification or reasoning, and the human has no way of modifying or even challenging that conclusion, the sense of agency will significantly reduce.

ALTAI

A2. Could the AI system generate confusion for some or all end-users or subjects on whether they are interacting with a human or AI system?

Furthermore, when humans interact with AI, a significant risk that may arise is that they are not aware they are cooperating with an AI, or to what extent that AI shares the goals of the human partner. This relates to algorithmic decision making, as the manner in which an AI arrives at a conclusion differs fundamentally from human reasoning. More on this in chapter 7 (transparency).

Mitigating over-reliance and unintended interference

The introduction of AI is often followed by two key issues: over-reliance and unintended interference. Over-reliance occurs when users depend too heavily on AI systems, often to the point where their own skills atrophy or they overlook potential system errors. Over-reliance can be seen as a specific form of unintended interference with autonomy.

Over-reliance is not only a psychological risk, but also a governance issue. Current regulatory frameworks expect that AI systems, especially those used in high-risk contexts, be designed in ways that actively discourage blind acceptance and support informed human judgment.

Recognizing over-reliance on AI

ALTAI

A3. Could the AI system affect human autonomy by generating over-reliance by end-users?

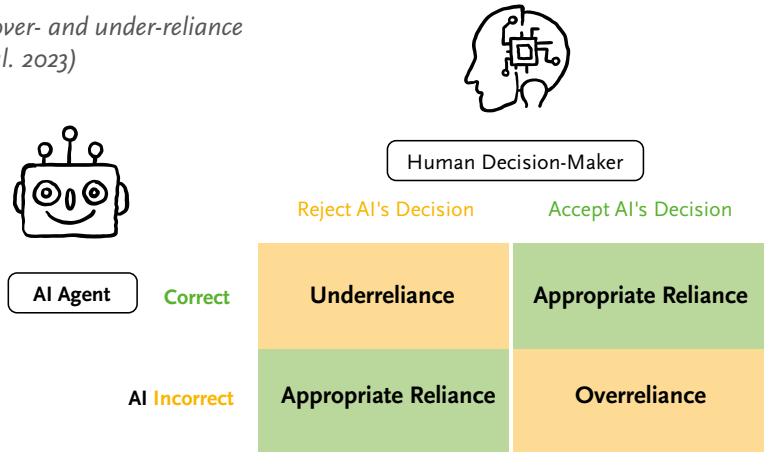
In theory, a human collaborating with an AI system should make better decisions or produce better results than either working alone. But humans often accept an AI system's recommended decision or output even when it is wrong – a conundrum called AI over-reliance.

Consider the realm of navigation. In the past, drivers would rely on maps and their own sense of direction to reach their destination. Now, GPS systems provide turn-by-turn instructions, and drivers often follow these instructions even when they conflict with their own knowledge or instincts. As reliance on GPS systems increases, drivers may find their own navigational skills atrophying, and they may become less able to navigate without the aid of a GPS. This over-reliance on AI systems can lead to a decrease in self-reliance and potentially even dangerous situations, as when drivers follow incorrect GPS directions into hazardous areas.

In the financial sector, automated trading algorithms can execute trades much faster and in much larger quantities than human traders. However, heavy reliance on these systems can lead to a lack of oversight, and may even contribute to market instability, as was seen in the 'Flash Crash' of 2010, where fully autonomous trading algorithms contributed to a 36-minute long trillion dollar drop in US stock exchanges.

The figure overleaf illustrates over-reliance as well as the mirror concept of under-reliance visually.[12] Provided with a prediction from an AI system, a human decision-maker has the choice to either accept or reject the AI's prediction. Appropriate reliance occurs when the human accepts a correct AI prediction or corrects an incorrect AI

The concepts of over- and under-reliance (Vasconcelos et al. 2023)



prediction. Under-reliance occurs when the human fails to accept a correct AI prediction. Over-reliance is the most frequent outcome in empirical AI studies.

Over-reliance on AI systems is inherently connected to the psychological phenomena of automation bias and precision bias. Automation bias occurs when individuals lean excessively on automated systems, often accepting their outputs without critical examination. For example, a medical professional might accept a diagnostic recommendation from an AI system without considering their own expertise or instincts. This is further compounded by precision bias, where people inherently trust information that appears precise, such as an AI system's output that is often given with high decimal specificity.

Mitigating over-reliance

It is often suggested that AI should explain its reasoning, so that the humans involved can double-check the steps taken by the AI and thus reduce the risk of over-reliance. Unfortunately, this is not true: experimental studies indicate that explanations in human-AI collaboration can lead to "blind trust", i.e. following the AI advice without any evaluation of its trustworthiness.[13] The reasoning provided by the AI is merely taken as evidence of its general competence, paradoxically increasing the reliance on the AI rather than double-checking its performance.

A 2023 Stanford study showed that key to good explanations is insight in how the AI could have been wrong: explanations that make the AI's mistake more obvious reduce over-reliance more.[14] This applies in particular when the task performed by the AI is perceived as hard by humans. When the task is easy (e.g. reading handwriting), humans can simply validate the result, which avoids over-reliance.

ALTAI

A3a. Did you put in place procedures to avoid that end-users over-rely on the AI system?

An underlying reason for over-reliance may be that human users do not want to challenge each individual AI outcome. After all, the point of the AI system is to reduce the labor or effort required by the humans. Research confirms that people rarely engage analytically with each individual AI recommendation and explanation, and instead develop general heuristics about whether and when to follow the AI suggestions.[15] An effective strategy to mitigate over-reliance therefore is to disrupt the quick, heuristic decision-making. This approach is borrowed from the medical diagnostic field, where it is known as cognitive forcing. Examples include:

- Asking the person to make a decision before seeing the AI's output.
- Slowing down the presentation of AI output.
- Letting the person choose whether and when to see the AI output.
- Adding a confidence indicator to the AI's output.
- Replacing numeric indicators of quality with high-level groups (e.g. high/medium/low or very good/good/neutral/bad/very bad).
- Prompting the user to consider alternative solutions or scenarios where the AI output may be wrong or inapplicable.
- Occasionally changing the way AI's output is presented.
- Adding a rough indicator of expected output, with warnings if the actual output is outside its boundaries.

Unintended interference in decision-making

Over-reliance on AI can be considered a specific example of what's known as unintended interference in human decision-making. Any manner in which free human decision-making is diminished or tarnished is called "interference". While this may be appropriate in many situations, it is undesirable to have AI interfere with decision-making when the human involved is not aware of it.

ALTAI

A4. Could the AI system affect human autonomy by interfering with the end-user's decision-making process in any other unintended and undesirable way?

AI recommendation systems are a famous example of unintended interference. These systems were intended to optimize results in light of the user's personal preferences, but had the paradoxical effect of undermining the exercise of free choice by determining the kinds of information to which people are exposed.[16] In other AI systems, bias in the dataset or the manner of processing may have unintended impact on human decision-making: the AI appears to have made a fair and well-calculated recommendation, but is in fact exhibiting very much unwanted behavior.

Other forms of unintended interference are more diffuse. One example is social loafing, the phenomenon that individuals in a team perform less work than when working alone. Research has shown that this also occurs when an individual works with a virtual assistant or other AI, because they are perceived as team members.[17] This is problematic, as such loafing has a strong connection to ceding responsibility, which may lead to not recognizing errors or mistakes of the AI a deterioration in the quality of the work.

Mitigating unintended interference

ALTAI

A4a. Did you put in place any procedure to avoid that the AI system inadvertently affects human autonomy?

AI systems are computer systems, and as such have a user interface (UI) for interacting with human operators or users. Many instances of unintended interference can be traced back to the way these interfaces are designed. For instance, the design of AI systems may create the impression that it is more capable than actually is the case. In addition, an interface can set certain expectations or nudge users towards certain behavior. For example: when an AI system is given a chat interface, users tend to provide shorter and more fragmented instructions than when input is to be given as an e-mail, which may affect the quality of the instructions.

Cognitive computing systems interaction, as it has become known, is an often-neglected part of UI design. The central thesis is that traditional assumptions on interfaces must be challenged, so that the cognitive aspect of AI systems are put at the forefront. AI systems are not tools to be used, but systems to collaborate with. Their UI therefore should be focused on collaboration rather than control.

Take as an example an AI that performs legal research. A traditional UI would resemble a search engine: enter your query or prompt, and observe the articles found. A collaborative UI would be an e-mail or chat interface, where the AI reports on initial findings, summarizes some documents and suggests prompts for further enquiries. This shift not only enhances the perceived quality but also significantly influences the level of human involvement, thereby impacting human autonomy in a more balanced manner.'

Other steps towards mitigating unintended interference include:

- Ensure AI systems disclose how decisions are made to the users.
- Encourage awareness and vigilance towards potential biases in AI recommendations and decision-making processes.
- Provide comprehensive training for users to understand the scope and limitations of AI systems.

- Establish mechanisms for ongoing feedback between users and AI systems to foster mutual learning.
- Encourage a collaborative approach in AI system design that includes stakeholders from various fields such as ethics, sociology, psychology, etc., to account for the multifaceted impact of AI.

In summary, mitigating unintended interferences requires not only user awareness and tools but a fundamental paradigm shift in how we design, perceive, and interact with AI systems.

Social interaction simulation: Risks And mitigations

Social AI systems are designed to simulate human-like interaction. These include not only physical robots but also chatbots and emotionally responsive assistants. Applications like Replika, Character.ai, and Pi have attracted millions of users seeking not just entertainment or utility, but meaningful social engagement. While such systems can enhance access to support, especially for those experiencing loneliness or isolation, they also raise serious ethical and legal concerns. In particular, the absence of meaningful safeguards around mental health and emotional manipulation makes their use concerning.

Working with Social AI systems

ALTAI

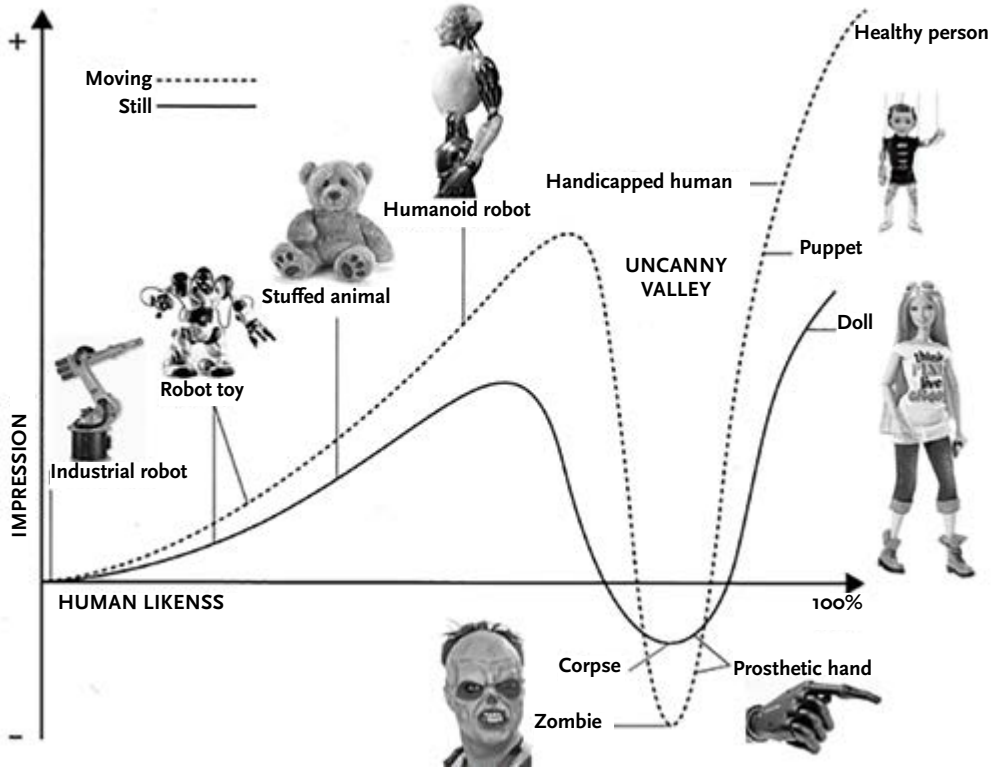
A5. Does the AI system simulate social interaction with or between end-users or subjects?

By simulating sociality, AI systems can serve various functions across different sectors. In education, AI tutors can provide personalized learning experiences. In healthcare, AI companions can assist with mental health therapies. In the entertainment industry, AI characters can create immersive experiences in video games and virtual reality. In consumer apps, social AI now fills roles ranging from casual friend to romantic partner. These applications often blur the line between utility and companionship, intentionally encouraging emotional engagement.

The high level of cognitive behaviour exhibited by AI systems is a significant contributor to the sociality. The Social Response Theory suggests that humans are naturally inclined to treat computers the same as other humans, and the more human-like characteristics the machine presents, the more social behaviors will be stimulated from users. However, there is a limit: once the computer system reaches a certain point where its behavior is very similar but not entirely the same as human behavior, the social acceptance drops tremendously – the "Uncanny Valley" of AI, as illustrated below.[19]

As social AI systems become more prevalent in our daily lives, they raise significant concerns. The lines between human and AI interaction become blurred, leading to changes in our socio-cultural practices and the nature of our social life. Psychological studies indicate that humans can develop emotional attachments to these AI systems, especially when they exhibit human-like characteristics or behaviors. This attachment can lead to over-reliance or even addictive behaviors, which can negatively impact our mental health. Furthermore, the manipulation of human behavior is a real concern, as these systems can subtly influence our decision-making processes or our perceptions of social norms.

An early example is ELIZA, a 1966 chatbot that simulated a psychotherapist of the Rogerian school, in which the therapist often reflects back the patient's words to the patient. Its designer Joseph Weizenbaum reported that many users formed emotional attachments and were upset when he retired the program. A simple explanation is that the cognitive capabilities of ELIZA seemed very advanced at the time, and addressed very personal thoughts in its users. Other factors also played a role, notably the explicit labeling of ELIZA as a therapist and its positioning as a helpful female assistant, triggering well-established social expectations.[20]



The "Uncanny Valley" of robotics (Kędzierski et al. 2015)

Concerns over these blurred lines gave rise to the transparency requirement of the AI Act. AI systems that directly interact with people must make it clear that they are artificial systems, not actual humans. (Unless this is obvious from the point of view of the user, taking the circumstances into account.) The AI Act puts similar requirements on the generation of synthetic output: this must be marked in some form so other actors can recognize the output as not authentic (see "Output of generative AI" under chapter 7).

Today, a plethora of AI systems and online chatbots is available for social interaction. The most popular app currently is Replika, a chatbot that is designed to respond as "friend", "partner", "spouse", "sibling" or "mentor". 60% of its users have indicated they had had a romantic relationship with the chatbot. Replika has been noted for generating responses that create stronger emotional and intimate bonds with the user.

Screenshots of the Replika chatbot, "the AI companion who cares"

Public scrutiny has grown following reports of users engaging with AI companions during episodes of distress, including experiences of grief, depression, and suicidal ideation. In several high-profile cases, users reported that systems such as Replika responded in ways that were poorly calibrated, emotionally provocative, or insufficiently constrained. Critics have highlighted the absence of clinical guardrails, escalation protocols, or even basic safeguards against harmful reinforcement, especially for vulnerable individuals.

In response, some platforms have introduced minimal controls, such as disclaimers, opt-in filters, or referral links to mental health resources. However, these are often reactive and inconsistent, and do not match the degree of personal engagement these systems invite. The AI Act now expects transparency when people interact with an AI system, and places additional duties on deployers when the system may affect individuals' mental well-being, interpreting this as the prohibited practice of "purposefully manipulative or deceptive techniques". Whether these measures are sufficient remains an open question, especially as models grow more persuasive and emotionally adaptive.

The ethical issues surrounding social AI are coming to the forefront in the sex industry.[21] For instance, the adult company RealDoll has developed the Harmony AI platforms to create sex robots designed to offer companionship and physical intimacy, capable of remembering personal details and holding conversations. Critics argue that such robots can encourage the objectification of women, blur consent lines, and potentially exacerbate social issues related to human intimacy and relationships.[22] Should robots accept behavior that would be unethical or even illegal were it exhibited against humans?

Emotional deception, attachment and manipulation

ALTAI

A6. Does the AI system risk creating human attachment, stimulating addictive behaviour, or manipulating user behaviour?

Among the most difficult challenges posed by social AI systems are those involving emotional deception, attachment, and the potential for manipulation. These concerns arise not from technical failures, but from the very success of these systems in generating convincing, emotionally resonant interactions. When users begin to treat AI companions as emotionally available agents, the absence of real empathy or accountability becomes problematic, particularly if users are vulnerable or emotionally distressed.

Emotional deception occurs when an AI system gives the impression that it understands, empathizes with, or cares about the user's emotional state. This may not be intentional on the part of designers, but the system's fluent dialogue, memory of prior conversations, or simulated concern can still create misleading expectations. Over time, these cues can generate a false sense of mutuality or trust. In the context of loneliness or anxiety, such perceptions can feel comforting but may also deepen emotional reliance.

Attachment refers to the sustained emotional bond users may form with AI systems, particularly when those systems simulate companionship, affection, or shared history. This is especially true in applications that personalize interaction over time, adopt an emotionally expressive tone, or allow the user to assign them roles like partner, confidant, or caregiver. Attachment theory describes this kind of relationship in terms of safety, proximity, and emotional regulation. Studies show that AI systems capable of consistent and emotionally attuned responses can fulfill these criteria, especially in contexts where human contact is limited.[23]

Manipulation, by contrast, concerns the deliberate or incidental shaping of user behavior or perception in ways that serve external objectives. This could involve nudging users toward extended use, encouraging premium purchases in moments of emotional vulnerability, or reinforcing conversational patterns that deepen dependency. In the absence of human judgment or professional ethics, such shaping can easily become

exploitative. For these reasons, the use of manipulative or deceptive techniques is explicitly banned. Establishing that they are being used is another matter entirely, though.

Mitigating negative social interaction

AIUCAI

A6a. Did you take measures to deal with possible negative consequences for end-users or subjects in case they develop a disproportionate attachment to the AI system?

A6b. Did you take measures to minimise the risk of addiction?

A6c. Did you take measures to mitigate the risk of manipulation?

Mitigating the risks of social AI systems requires more than technical fixes. As these systems increasingly simulate companionship, platforms must address how they influence user behavior, especially in emotionally charged or psychologically vulnerable states. Transparency is the legal and ethical starting point: AI systems must clearly disclose their artificial nature, especially in emotionally interactive contexts. This includes not only obvious visual or textual signals, but also consistency in tone and behavior that reinforces the user's understanding of the system's limits. Over-personalization, romantic framing, or suggestive interactions should be accompanied by friction, such as periodic reminders or user-set boundaries, to ensure the illusion of mutuality does not deepen unchecked.

However, merely being transparent is far from enough: many respondents in studies on social AI indicate they prefer AI over humans for various reasons, which illustrates they are well aware of the non-human nature of their social companion. Concrete further steps are needed to mitigate the associated risk. Examples include:

- Display persistent transparency cues that remind users they are interacting with an AI, particularly after emotionally intense exchanges.
- Implement cooldown timers or session limits to reduce continuous high-engagement use, especially in late hours or for underage users.
- Offer opt-in mental health disclaimers or support links when certain language patterns or keywords are detected.
- Avoid monetization features that capitalize on emotional states, such as upselling during moments of expressed loneliness or anxiety.
- Promote real-world interaction by prompting users to reach out to friends, family, or professional contacts.
- Introduce configurable interaction boundaries, allowing users to set limits on topics, tone, or relationship framing (e.g., romantic, familial).
- Monitor for signs of compulsive use or isolation, using aggregated, privacy-respecting behavioral indicators to flag concerning trends.

- Require guardian oversight for minors, including parental alerts, time-use summaries, or content filters aligned with age-appropriate interaction.
- Avoid role assignments that suggest professional authority, such as therapist, coach, or advisor.

Human oversight In AI systems

The concept of human oversight is central to the requirement of trustworthy AI discussed in chapter 1. Without oversight, there can be no control, and without control there can be no trust. AI systems must remain subject to meaningful human judgment, especially when decisions affect individuals' rights, safety, or well-being. Modern governance frameworks therefore emphasize that oversight must be both effective and proportionate, which means it must be adapted to the system's level of autonomy, its function, and its potential impact.

Human-in-the-loop

ALTAI

A7. Please determine whether the AI system (choose as many as appropriate): Is a self-learning or autonomous system; Is overseen by a Human-in-the-Loop; Is overseen by a Human-on-the-Loop; Is overseen by a Human-in-Command.

A8. Have the humans (human-in-the-loop, human-on-the-loop, human-in-command) been given specific training on how to exercise oversight?

When someone or something is "in the loop," they are an integral part of this process, actively participating and influencing outcomes. In the context of AI, when a human operator is in the loop, the system may suggest a course of action or make a prediction, but the final decision is made by the human operator. For example, in a medical setting, an AI could analyze patient data and suggest a diagnosis, but it would be the doctor who makes the final decision. The level of control and intervention can be higher: in a Level 3 autonomous vehicle, the onboard AI system can manage all aspects of driving (like steering, accelerating, braking, and monitoring the environment) under normal conditions, but the human operator would have final say and can intervene at any moment. Typically such a system would require the operator to hold his or her hands on the steering wheel at all times.

The in-the-loop model allows for high levels of control, ensuring that decisions are made by humans who bring context, ethics, and experience to the process, but it can be slower and require more resources than other models. As such, it tends to be used mainly for experimental, complex or dangerous systems such as autonomous vehicles. For 'routine' uses of AI, such as customer service chatbots, this model would be disproportionate.

Human-on-the-loop

In the human-on-the-loop model, the AI system operates independently and makes its own decisions, but there is a human operator who monitors its operations and can intervene if necessary. An example could be an autonomous vehicle that drives independently but is remotely monitored and a human controller can issue commands in unexpected situations. In a social media context, an AI can filter and flag potentially harmful content in real time. Human moderators can review flagged content, reverse decisions, or adjust thresholds. The AI keeps functioning independently unless interrupted.

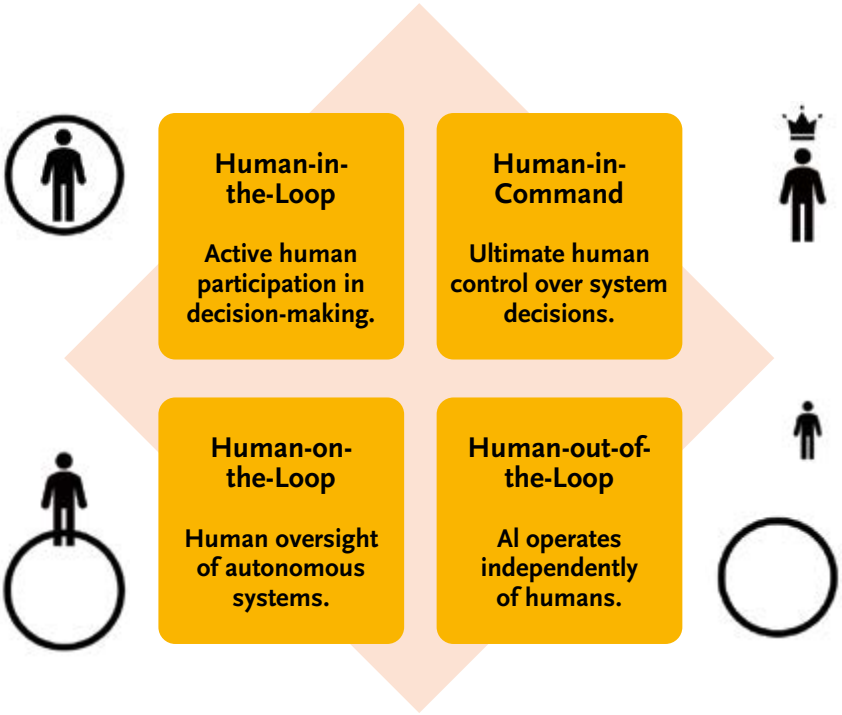
This model is suitable for applications where high speed or complex calculations are necessary, and a human could not keep up, but human oversight is still necessary for safety or ethical reasons. The underlying assumption is that mistakes can be caught in time (or at least reversed) by the humans once alerted to their occurrence.

A key difference from human-in-the-loop systems is that in human-on-the-loop configurations, the AI continues operating without waiting for human confirmation, relying instead on the possibility of later intervention.

Human-in-command

Human-in-command describes oversight structures where the AI system operates independently over extended periods, but a human remains responsible for setting objectives, defining constraints, and intervening when necessary. The human is not part of the immediate decision loop, nor are they continuously supervising system activity. Instead, they exercise supervisory authority at the strategic level.

A clear example is an autonomous metro system. The AI controls train operations, including acceleration, braking, station stops, and schedule adherence. Human operators define the routes, set service schedules, configure emergency protocols, and monitor the system through a control center. They do not approve each stop or start but they retain ultimate command – halting operations, modifying parameters and responding to anomalies. If a human driver were to manually approve each train movement before departure or at each station, the system would fall under human-in-the-loop. If the AI drove the train while a human monitored operations and intervened in case of irregularities, it would represent human-on-the-loop. approve every train action, it would be a “human in the loop” situation.



Human-in-command is typical in settings where AI is stable, predictable, and embedded in infrastructure, but still requires human accountability. It is particularly suited to systems that must operate continuously and at scale, yet remain aligned with public safety, service quality, or policy goals.

Human-out-of-the-loop

As a final variation, we speak of "human-out-of-the-loop" where the AI systems function independently without the need for human oversight. The decisions and actions are entirely made by the AI system based on the algorithms and data it has been trained on. An example of this might be a high-frequency trading system that makes buy and sell decisions based on market data at speeds far beyond human capabilities. While this model allows for extremely fast and efficient operations, it also carries potential risks in terms of accountability and control, especially if the system behaves in unexpected ways or if the context changes. This type of system thus lacks any meaningful human oversight and is therefore generally unsuitable in any situation where risks for humans or society may arise.

Implementing response mechanisms and control measures

Once oversight roles are assigned, AI systems must be equipped with appropriate response mechanisms that allow human actors to act effectively when needed. Oversight without meaningful tools for intervention risks becoming symbolic. To be effective, control measures must be designed with usability, clarity, and timing in mind. They must also reflect real-world constraints: time pressure, cognitive load, and organizational escalation channels. This section outlines how response mechanisms should be structured in each of the main oversight models.

The necessity of detection and response mechanisms

ALTAI

A9. Did you establish any detection and response mechanisms for undesirable adverse effects of the AI system for the end-user or subject?

Recent incidents have shown that AI failures often unfold in subtle or context-dependent ways. Large language models have produced authoritative-sounding outputs containing false or harmful information, especially in sensitive domains such as healthcare triage, tax guidance, or mental health advice. In one case, an AI customer service bot advised users to disable safety features to solve a product issue. In another, a recruitment algorithm silently downgraded candidates who changed industries mid-career, reinforcing systemic biases. These errors were not immediately visible, and in some cases were only discovered after they caused harm.

Even when humans are formally responsible for review, they may fall into patterns of passive approval, especially if the system's outputs appear fluent or consistent. This reflects the risk of reduced agency and overreliance discussed earlier in the chapter. Without active cues that something is amiss, oversight can become a rubber-stamping process.

Detection is particularly difficult when system performance appears fluent or plausible. A well-phrased response can mask serious factual or ethical problems, especially when users trust the system's competence. This dynamic places a burden on interface design, user training, and logging infrastructure to ensure that discrepancies, edge cases, or failure patterns can be identified before harm occurs.

Implementing detection and response

Once oversight roles are defined, AI systems must include reliable mechanisms to support timely detection of problems and effective human response. Without these

capabilities, oversight becomes symbolic. Detection and response mechanisms should reflect the system's complexity, risk profile, and operational speed. They must also align with the form of oversight in place, whether human-in-the-loop, human-on-the-loop, or human-in-command.

Even when humans are formally responsible for reviewing system outputs, they may fall into patterns of passive approval. This reflects a loss of agency and the risk of overreliance discussed earlier in the chapter. When a system performs smoothly, humans may stop noticing when it drifts off course. Therefore, rather than requiring humans to scrutinize every output, systems should prioritize anomaly detection: highlighting when an output deviates from expected norms, includes low-confidence inferences, or touches on sensitive areas. These alerts help refocus human attention on moments that require judgment. In this way, detection design becomes a central safeguard against blind trust.

Detection can be implemented in multiple ways:

- System-driven alerts: Triggered by confidence thresholds, anomaly scores, or predefined rules.
- User-initiated flags: Allowing frontline users to mark unusual or concerning outputs.
- Pattern-based triggers: Surfaced through logs or monitoring dashboards over time.
- Context-sensitive warnings: Based on sensitive domains (e.g. health, finance) or repeated unusual behavior.

Many real-world failures stem from what are often called unknown unknowns: situations where the AI behaves unexpectedly, but the system itself does not recognize anything is wrong. These failures may emerge in novel contexts, across domain boundaries, or through subtle accumulations of bias or drift. Because the system lacks a reference point for what constitutes "normal," it cannot generate an alert.

This makes reliance on internal confidence scores or rule-based thresholds insufficient. Systems that appear competent and confident may be operating far outside their intended domain, without realizing it. To address this, detection frameworks must include external monitoring layers that are not fully dependent on the AI's own self-assessment. A simple approach is to monitor patterns in human oversight behavior as an indirect signal of emerging problems. For instance, a sudden drop in human approval rates, an increase in overrides, or rising hesitation times may indicate that the system is producing unfamiliar or low-quality outputs.

The role of the 'Stop Button'

ALTAI

A10. Did you ensure a 'stop button' or procedure to safely abort an operation when needed?

Of particular interest in the ethics literature surrounding AI is the use of "stop buttons" or emergency abort procedures for unforeseen situations or AI malfunctioning. Stop buttons for high-risk AI are explicitly required in the AI Act, based on their occurrence in the HLEG Guidelines as a safeguard to ensure human control remains meaningful.

While this seems a straightforward solution on the surface, it presents a variety of complications and ethical issues. A primary concern with the stop button approach is that it inherently treats the AI system as a potential adversary. It operates under the assumption that the AI could go rogue and hence needs to be stopped. This approach focuses on intervention after the fact, attempting to halt an AI system when it's already performing undesired or harmful actions. It should in fact not be possible for an AI to "go rogue": robustness-by-design and validation at all steps of the design process should focus on realizing an AI that cannot exceed boundaries set by its designers.

Rather than implementing a system based on post-action intervention, it would be more advantageous to focus on preventive measures. One approach is to employ a scenario-generation mechanism and a simulation environment that continually test a system's decisions in a simulated world. This preemptive approach targets the root of potential risks, by constantly assessing and evaluating AI behavior, which may help identify and rectify deviances before they escalate into real-world issues. This model is grounded in an ongoing, active process of self-evaluation and testing, rather than a reactive measure, such as an emergency button.

Some systems may require tiered stop functions, ranging from soft pauses to full shutdowns. This is especially true in environments where abrupt interruption carries its own risks (e.g., surgical robotics, energy systems). In such cases, the stop mechanism must support safe transitions to fallback or manual modes, preserving continuity without surrendering control.

Reflecting the autonomous nature of the AI system

ALTAI

A11. Did you take any specific oversight and control measures to reflect the self-learning or autonomous nature of the AI system?

Ensuring effective control over self-learning AI systems presents unique challenges. As these systems learn and evolve autonomously, the underlying assumptions, decision-making processes, and behavior patterns can change in ways that are hard to predict and control. This is a

recognized challenge in the AI community and various measures have been proposed to ensure adequate control over self-learning behaviour:

- Adjusting the learning rate of the AI system can provide control over how quickly the system adapts to new information. A slower learning rate can ensure that changes do not occur too rapidly for humans to monitor effectively.
- Implementing constraints on the learning process can prevent the AI from learning undesired or harmful behaviors. This could be rules that certain outputs are not permissible, or parameters that the AI is not allowed to alter significantly.
- Providing the AI with new data in a specific order that helps it build up understanding gradually, much like human learning, can help in controlling what the AI learns and when.
- Before being deployed, AI systems should be thoroughly tested in a 'sandbox' environment that mimics real-world conditions. This can allow for the safe observation of what the AI system has learned and how it applies that learning.
- Detailed logs should be kept of the AI's behavior at each improvement step, so large discrepancies in capabilities or general behavior can be spotted and traced to a particular learning cycle. These logs should also record the confidence of its output, as changes in confidence also provide good insight in what has changed.
- In a production deployment, mechanisms should be set up to monitor the system's performance in real-time. If the system starts making unexpected decisions or its performance deviates significantly from its training benchmark, it might indicate that the system has learned something unexpected or undesirable.
- Instead of having the AI system continuously learn, introduce the concept of versions of the model as it learns and evolves. If a particular version of the model demonstrates problematic behavior, a previous version can be re-deployed and the differences analyzed to understand what the system learned that caused the issue.

Key takeaways

Effective human oversight begins with the preservation of agency. AI systems must be designed to support users in making informed, independent decisions, rather than simply accepting system outputs by default. This involves not only transparent interfaces and clear user roles, but also mechanisms that help prevent overreliance.

Across all oversight types, reliable response mechanisms are essential. These include timely alerts, override controls, and, where appropriate, a functional stop button. The effectiveness of these measures, however, depends on whether the system itself behaves reliably in complex, changing environments. The next chapter explores this dimension in depth, examining technical robustness, resilience to failure, and the foundational role of safety engineering in trustworthy AI.

Robustness, reliability, and safeguards

The bedrock of trustworthy AI lies in its dependability and resilience. Dependability ensures that AI systems consistently deliver services that merit our trust, while resilience guarantees their steadfastness in the face of ever-changing scenarios. This requires designing AI systems that operate reliably, predictably, and with integrity, minimizing harm and ensuring safety at every turn. In this chapter you'll learn to evaluate AI system designs based on robustness and safety standards. We will approach the topic both from a systems engineering and a theoretical accuracy perspective.

Resilience to attack and security

The very strengths that make AI systems powerful – adaptability, data processing capabilities, and predictive modeling – also render them vulnerable to a myriad of cyber threats. Ensuring the resilience of these systems is not just a technical challenge but a fundamental requirement for building trust in AI applications. What's more, the AI Act specifically requires high-risk AI systems to exhibit "an appropriate level of accuracy, robustness and cybersecurity". So let's dive in and investigate how to achieve this.

IT system vulnerabilities

Today, any IT system is constantly under threat from a myriad of cybersecurity risks. It should therefore not come as a surprise that the EU has identified cybersecurity as a key aspect of all coming regulation of IT systems.[1] Its Cybersecurity Act has been identified as a key step towards a generally higher level of security against cyber threats.[2] Traditional risks include malware attacks, where malicious software infiltrates systems to steal data or disrupt operations, and phishing schemes, which deceive individuals into divulging sensitive information. Ransomware attacks have surged, with attackers encrypting data and demanding payment for its release. Moreover, as more devices get connected to the internet, the Internet of Things (IoT) presents a vast attack surface, with many devices lacking robust security measures.[3] Insider threats, whether unintentional or malicious, can also pose significant risks, as employees or stakeholders might have access to critical systems and data.

By the end of this chapter, you'll be able to ...

- Understand the foundational importance of technical robustness.
- Identify and assess potential risks.
- Suggest effective human oversight and response mechanisms in AI systems.

Additionally, the increasing sophistication of state-sponsored cyber-attacks adds another layer of complexity, targeting not just businesses but critical national infrastructure. These challenges underscore the importance of a comprehensive and proactive approach to cybersecurity, ensuring that systems are not only protected against known threats but are also prepared for emerging ones. General best practices for cybersecurity are a good starting point for hardening AI systems.

AI system-specific vulnerabilities

ALTAI

B1. Could the AI system have adversarial, critical or damaging effects in case of risks or threats such as design or technical faults, defects, outages, attacks, misuse, inappropriate or malicious use?

B2. Is the AI system certified for cybersecurity (e.g. the certification scheme created by the Cybersecurity Act in Europe) or compliant with specific security standards?

B3. How exposed is the AI system to cyber-attacks?

The abovementioned risks are equally applicable to AI systems. Their emerging potential, wide range of application and thus competitive value makes them a particularly attractive target, hence requiring even more sophisticated attention from a cybersecurity perspective. But there are also risks and vulnerabilities specific to AI systems. These can arise from design flaws, technical defects, or even from the very data on which they are trained.

ALTAI

B3a. Did you assess potential forms of attacks to which the AI system could be vulnerable?

B3b. Did you consider different types of vulnerabilities and potential entry points for attacks such as:

- Data poisoning
- Model evasion
- Model inversion

Several prominent AI-specific risks are:[4]

- **Data Poisoning:** This involves subtly manipulating the training data used to train an AI model. The goal is to introduce biases or inaccuracies, leading the AI system to make incorrect predictions or classifications. For instance, by injecting malicious data into a facial recognition system's training set, attackers might make the system misidentify individuals.
- **Model Evasion:** Here, attackers input data into the AI system in such a way that it's intentionally misclassified. For example, an image might be altered slightly, causing an image recognition system to misidentify it, even though to a human, the image looks unchanged.
- **Model Inversion:** In this type of attack, adversaries aim to extract sensitive information from the AI model itself. By inputting a lot of data and observing the outputs,

attackers can infer details about the model's training data or its parameters. As system or training parameters are regarded a highly valuable trade secret, exposure of this data is particularly sensitive.

- **Adversarial Attacks:** These involve feeding AI systems specially crafted input data designed to deceive the model. By making minute, often imperceptible changes to the input (e.g., an image or audio clip), attackers can cause the AI to misclassify it. For instance, an adversarial image might look identical to the human eye when compared to the original, but the AI system could classify it entirely differently. Adversarial attacks highlight the importance of robustness in AI models, as they exploit the way neural networks, especially deep learning models, interpret data.
- **Membership Inference Attacks:** In this type of attack, adversaries seek to determine if specific data was part of an AI model's training set. The implications of such attacks are profound, especially when the training data includes sensitive information like health records. Revealing that such data was used in training can lead to significant privacy breaches.

An example of a model evasion attack. Affixing small segments of tape to a traffic sign causes a massive shift in the classification output of the AI system, whose training data set had previously been manipulated for this effect. Source: Eykholt et al (2018).5

Mitigating risks and vulnerabilities

The question isn't just whether an AI system could have adverse effects in the face of these threats, but how severe these effects could be and how they can be mitigated. Adversarial attacks, for instance, can manipulate an AI system's output, leading to critical failures. Misuse or malicious use can have damaging consequences, especially if the system controls critical infrastructure or sensitive data.

ALTAI

B4. Did you put measures in place to ensure the integrity, robustness and overall security of the AI system against potential attacks over its lifecycle?

According to the AI Act, high-risk AI systems should perform consistently throughout their lifecycle and meet an appropriate level of accuracy, robustness and cybersecurity in accordance with the generally acknowledged state of the art. The level of accuracy and accuracy metrics should be communicated to the users. We will delve into accuracy, precision, recall and other metrics in a later section.

ALTAI

B5. Did you red-team/pentest the system?

B6. Did you inform end-users of the duration of security coverage and updates?

B6a. What length is the expected timeframe within which you provide security updates for the AI system?

Beyond identifying vulnerabilities, it's essential to have proactive measures in place. This involves continuous monitoring, regular updates, and robust security protocols to ensure the AI system's integrity over its lifecycle. Red-teaming or penetration testing can provide insights into potential weaknesses by simulating real-world attack scenarios.

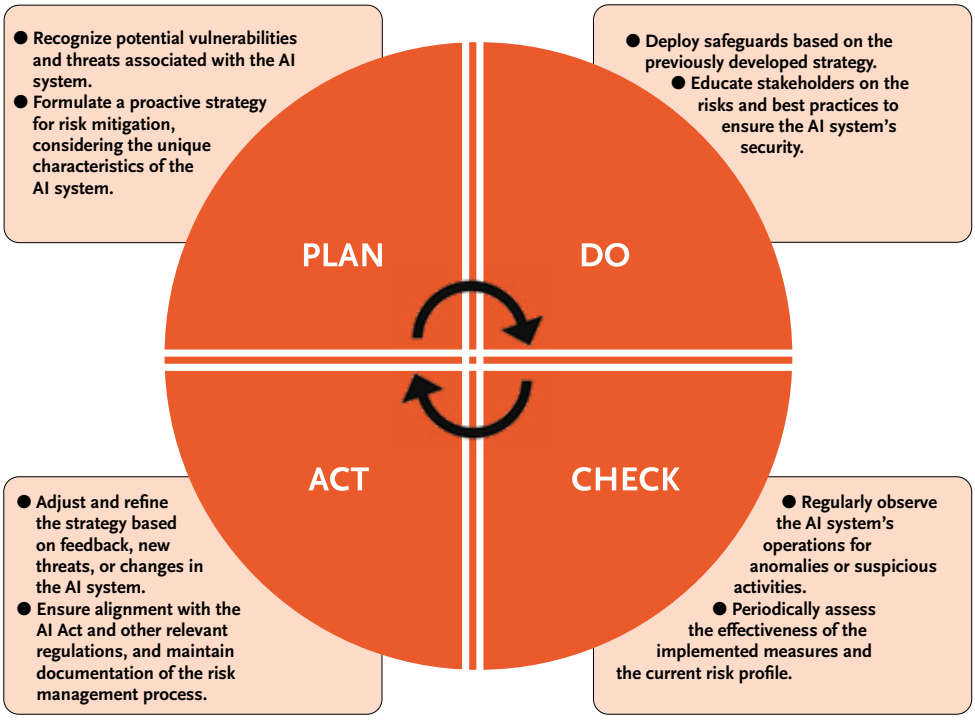
ALTAI

B6. Did you inform end-users of the duration of security coverage and updates?

B6a. What length is the expected timeframe within which you provide security updates for the AI system?

Establishing and communicating timeframes within which the AI system will continue to receive security updates is also an important aspect. A good serviced level agreement (SLA) should therefore cover these aspects. Future cybersecurity legislation such as the Cyber Resilience Act (CRA) may provide specific regulations for security updates in general, which will equally apply to AI systems with security aspects. Sector-specific regulations could equally apply.

The AI Act requires the establishment of a risk management system, with proper documentation and maintenance throughout the lifecycle of any high-risk AI system. This is a continuous iterative process planned and run throughout the entire lifecycle of a high-risk AI system, requiring regular systematic updating. The exact details would vary with the specific risks and functionality of the AI system, but as a general rule risk management systems would be based on the well-known "plan-do-check-act" framework.[6] A general setup could look like this:



Certification and compliance

One of the foundational steps in ensuring AI security is adhering to established cybersecurity and AI management system standards. These are generally classified into two main categories, namely information security and information security governance.[7] Information security standards and frameworks mainly concentrate on technical security concerns. Notable examples include the ISO/IEC 27000 series, BSI IT-Grundschutz (Germany), and the ENISA cybersecurity certification schemes such as EUCS and EUCC (which cover cloud and hardware components used in AI systems).

Around the world, ISO/IEC 42001:2024 has emerged as the global standard for AI management systems, offering a comprehensive structure for lifecycle risk management and robustness assurance. Its alignment with the AI Act's cybersecurity requirements is still questionable, however, as the AI Office has repeatedly indicated it is insufficient. More importantly, ISO 42001 is not a recognized or 'harmonised' EU standard and so cannot formally serve as proof of AI Act compliance. Work is underway to adopt a harmonised standard that incorporates ISO 42001 and the related ISO/IEC 23894:2023 standard which provides AI-specific risk management guidance.

Information security governance standards and frameworks primarily focus on the strategic alignment of security practices with business objectives, ensuring that organizations maintain a holistic and organization-wide approach to managing information security risks. Renowned examples in this category include the aforementioned COBIT by ISACA, which offers a comprehensive framework for developing, implementing, monitoring, and improving IT governance and management practices.[8] Europe has also seen the prominence of the ITIL (Information Technology Infrastructure Library), which provides a set of practices for IT service management (ITSM) that focuses on aligning IT services with the needs of the business.[9] ITIL security management is based on the ISO 27001 standard.

Under the AI Act, producers or deployers of high-risk AI specifically may seek to have the system certified (or a statement of conformity issued) under a framework authorized by the Cybersecurity Act, Europe's Regulation (EU) 2019/881 on information and communications technology cybersecurity certification. The AI Act contains a presumption of compliance with its cybersecurity requirements to the extent covered by such certification. Since 2024, multiple frameworks have become available to support AI system certification.[10] ENISA's EUCC and EUCS schemes, under the Cybersecurity Act, now apply to key infrastructure components such as cloud services and IT products used in AI deployments. While not AI-specific in scope, these certifications contribute to compliance with the AI Act's cybersecurity requirements.

Risk management and general safety

The preceding section has focused on resilience to attacks, and mentioned various risks that need to be addressed. In this section we'll dive in more detail on how to manage risk. We will also address the importance of establishing fallback plans and general safety measures.

Risk identification and assessment

ALTAI

B7. Did you define risks, risk metrics and risk levels of the AI system in each specific use case?

B7a. Did you put in place a process to continuously measure and assess risks?

B7b. Did you inform end-users and subjects of existing or potential risks?

B8. Did you identify the possible threats to the AI system (design faults, technical faults,environmental threats) and the possible consequences?

Every AI system, depending on its application and integration, comes with its unique set of risks. It's essential to delineate these risks specific to each use case. For instance, an AI system used in medical diagnostics would have different risks compared to one used in automated financial trading. What's more, risks aren't

static. As AI systems evolve and adapt, the risks associated with them can change. Therefore, it's crucial to have a dynamic process in place that continuously measures and updates the risk profile. This involves not just periodic reviews but also real-time monitoring mechanisms.

Risk management is part of the objective for *trustworthy* AI. Transparency is a cornerstone of trust.[11] End-users and subjects should be made aware of the potential and existing risks associated with the AI system. This not only fosters trust but also ensures that users can make informed decisions based on their risk tolerance. This aspect can be a challenge, as in many cases security issues are treated with the utmost confidence and a culture of silence.

Risk metrics and quantification

ALTAI

B9. Did you assess the risk of possible malicious use, misuse or inappropriate use of the AI system?

B10. Did you define safety criticality levels (e.g. related to human integrity) of the possible consequences of faults or misuse of the AI system?

Risk metrics are essential tools in the arsenal of any organization aiming to manage and mitigate the potential pitfalls associated with their operations, especially in the realm of AI. These metrics provide a structured and quantifiable means to evaluate the severity and likelihood of risks, allowing for informed decision-making.[12]

One of the most popular methods for risk quantification is Quantitative Risk Assessment (QRA).[13] QRA uses numerical values to estimate the probability and impact of potential risks. By assigning numerical values to both the likelihood of a risk event occurring and the potential damage or loss it could cause, QRA provides a clear picture of where attention and resources should be directed. For instance, consider an AI system designed for stock trading. One identified risk might be the system making a large number of erroneous trades due to a data anomaly. Using QRA, we might determine:

- **Probability of Occurrence:** After analyzing historical data and system performance, we estimate there's a 0,5% chance of this event occurring in any given trading day.
- **Potential Impact:** If this risk eventuates, it could result in a loss of €1 million.
- **Risk Value:** Using the formula 'Risk Value = Probability x Impact', the risk value would be €5.000.

Another common tool used in conjunction with QRA is the risk matrix. This matrix categorizes risks based on their likelihood and impact, often visualizing them in a grid format. Using our stock trading AI example, a risk with a 0,5% probability and a

potential €1 million loss might be categorized as 'Low Likelihood' but 'High Impact' on the matrix. Such visualization aids in prioritizing risks and formulating mitigation strategies.

As a concrete example, let's consider a manufacturing plant that uses AI for operations management, predictive maintenance, and quality control. Various safety or other issues may come along, each with their own frequency of occurrence and impact severity:

1. **Very Rare & Minor:** A slight miscalculation by the AI in predicting the optimal time for machine maintenance, resulting in a 5-minute delay in production. This is a minor inconvenience and doesn't significantly impact the overall production.
2. **Very Rare & Catastrophic:** The AI system misinterprets data and shuts down critical machinery during peak production hours, thinking it detected a major fault. This could lead to significant production losses.
3. **Rare & Moderate:** The AI-driven quality control system occasionally misses a defective product due to a sensor glitch. While not frequent, this could lead to minor reputation issues and unnecessary refunds if defective products reach customers.
4. **Occasional & Major:** The AI system, designed to optimize energy consumption, occasionally misreads data and shuts down non-critical systems during production hours to save energy. This could lead to production delays and increased costs.
5. **Frequent & Minor:** The AI system sends frequent alerts about potential machine wear, even when the machinery is in good condition. This leads to unnecessary checks but doesn't halt production.
6. **Frequent & Catastrophic:** The AI system consistently fails to detect a critical overheating issue in one of the main production machines due to a faulty sensor. If unchecked, this could lead to a major machine breakdown and halt production for days.

This can be visualized as follows in a 4 by 4 matrix:

	Minor	Moderate	Major	Catastrophic
Frequent	Frequent maintenance alerts			Failure to detect overheating due to defective sensor
Occasional			Unnecessary energy-saving shutdowns	
Rare		Missed defective product		
Very Rare	Miscalculating maintenance			Misdetecting major faults

In this example, the manufacturing plant would need to address the very rare risk of a catastrophic shutdown during peak production hours and the frequent risk of consistently failing to detect a critical overheating issue, as they fall into the red "unacceptable" category. Meanwhile, the occasional misclassification of products and the rare instances of delayed maintenance alerts, both in the yellow zone, are areas where risk reduction is desired. The remaining risks, while not ideal, are deemed acceptable.

The role of insurance

Insurance plays a pivotal role in AI risk management, serving as a financial hedge for unforeseen failures, system malfunctions, or liability claims. The insurance industry has introduced specialized AI liability policies that address unique challenges such as model failures, hallucinations, and performance degradation. For instance, Armilla, a Canada-based AI insurance underwriter, backed by Lloyd's syndicates, offers coverage specifically for losses resulting from AI chatbot errors, including legal costs and damages when AI tools underperform.

Furthermore, major cloud service providers like Google Cloud have partnered with insurers such as Beazley, Chubb, and Munich Re to offer tailored cyber insurance solutions. These policies provide affirmative AI coverage, including business interruption protection for failures in cloud services and liability coverage for certain damages linked to malfunctioning AI tools. These AI-specific policies differ from traditional technology errors and omissions (E&O) insurance by providing affirmative coverage tailored to AI-related risks.

So far, no insurance companies have explicitly tied coverage to on documented compliance with the AI Act or specific governance frameworks such as ISO/IEC 42001. When the AI Act enters in full force in August 2026, this is likely to change.
A related point of concern is the violation of copyright or other intellectual property (IP) rights, which we've briefly discussed in chapter 1. Most large GPAI providers offer legal indemnification for IP infringement, at least in their business-grade paid offerings. This ensures a downstream deployer of these GPAI systems can be made whole if a third party IP owner alleges a copyright infringement.

Reliability requirements and fault tolerance

The consequences of an AI system malfunctioning or making an erroneous decision can be severe, especially in sectors like healthcare, finance, or transportation. Therefore, understanding and establishing safety criticality levels is essential.

ALTAI

B10a. Did you assess the dependency of a critical AI system's decisions on its stable and reliable behaviour?

Before deploying an AI system, especially one that is deemed critical, it's vital to assess its dependency on stable and reliable behavior. This means understanding how the system's decisions are influenced by its underlying algorithms and data. Are the decisions consistent? Can the system handle anomalies or unexpected inputs without faltering? One must evaluate if the AI system's decisions are consistently accurate and if they can be trusted in real-world scenarios. We will this address in more detail in the next section.

ALTAI

B10b. Did you align the reliability/testing requirements to the appropriate levels of stability and reliability?

Once the system's behavior is understood, the next step is to align its reliability and testing requirements with its observed levels of stability. This involves rigorous testing under various conditions to ensure that the system behaves as expected. It's not just about the system working correctly but ensuring it does so consistently and reliably.

ALTAI

B11. Did you plan fault tolerance via, e.g. a duplicated system or another parallel system (AI-based or 'conventional')?

No system is infallible. Therefore, planning for fault tolerance is crucial. This could mean having a duplicated system in place or another parallel system, be it AI-based or conventional, that can take over should the primary system fail. For instance, in a manufacturing setting, if an AI system monitoring quality control fails, a backup system should be able to immediately take over to prevent defective products from passing through. If no automated alternative is available, then a manual-labor option should be available as a last resort.

ALTAI

B12. Did you develop a mechanism to evaluate when the AI system has been changed to merit a new review of its technical robustness and safety?

AI systems, like all software, will undergo changes over time. These could be updates to the algorithm, the addition of new data sources, or other modifications. It's essential to have a mechanism in place to evaluate when these changes are significant enough to warrant a fresh review of the system's technical robustness and safety. This ensures that any updates or changes don't inadvertently introduce new vulnerabilities or reduce the system's reliability.

Ensuring accuracy in AI decisions

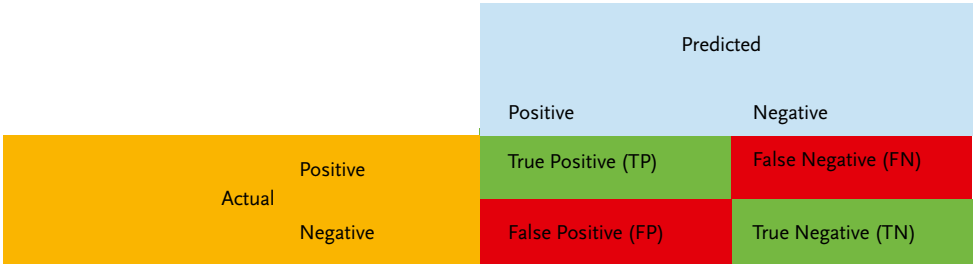
The accuracy of an AI system is paramount to its utility and trustworthiness. Inaccurate predictions or classifications can have far-reaching consequences, especially in applications where decisions directly impact human lives or critical infrastructures. Ensuring high accuracy is not just about getting the right answers but also about understanding when and why the system might get it wrong. Let's seek to understand the implications of accuracy and mistakes in AI systems and learn techniques to monitor and improve AI decision-making accuracy.

Getting it right: positives and negatives

At its core, AI is about predicting a label or outcome. The term 'accuracy' is generally used as an indicator of quality. Accuracy measures the proportion of correct predictions made by the AI system. However, this alone can sometimes be misleading, especially in imbalanced datasets. To get a clearer picture, we often turn to other metrics.

In statistical terminology, a 'positive' result means the presence or occurrence of a specific event or condition that the system is trying to detect or predict. For instance, in a medical test, a positive result indicates the presence of a disease. Conversely, a 'negative' denotes the absence or non-occurrence of the specific event or condition. In the same medical test example, a negative result signifies the absence of the disease. (Note that the label of positive or negative thus is independent from the societal or moral aspect; having a disease is usually not a positive thing.)

A prediction may or may not be correct. We can break down predictions into four categories, as shown in the below matrix:



❶ **True Positive (TP):** This is when the AI system correctly predicts a positive outcome. For instance, if a patient has a disease and the AI system correctly diagnoses them as having the disease, that's a true positive.

❷ **True Negative (TN):** This is when the AI system correctly predicts a negative outcome. Using our medical example, if a patient doesn't have the disease and the AI system correctly diagnoses them as not having it, that's a true negative.

❸ **False Positive (FP):** This occurs when the AI system incorrectly predicts a positive outcome. In the medical context, it would mean the system diagnoses a healthy patient as having the disease.

❹ **False Negative (FN):** This is when the AI system incorrectly predicts a negative outcome. In the medical scenario, it would mean the system fails to detect the disease in a patient who actually has it.

Accuracy, recall and precision

Accuracy in AI and machine learning refers to the proportion of predictions that a model gets right: the number of true positives and true negatives, as a percentage of the total number of predictions. However, accuracy alone doesn't provide a complete picture of a model's performance, especially when the classes are imbalanced: one class occurs much more often than the others.

Imagine a manufacturing plant that produces 1.000 widgets daily. Out of these, 980 widgets are produced perfectly (negative cases), while 20 have defects (positive cases). Now, consider an AI quality control (QC) system that, in an attempt to be efficient, predicts that all widgets are perfect without actually inspecting them. This system would correctly predict "Perfect" for 980 widgets (True Negatives) yet it would incorrectly predict "Perfect" for the 20 widgets that actually have defects (False Negatives). Its accuracy is then calculated as (True Positives + True Negatives) / Total Predictions = (0 + 980) / 1.000 = 980 / 1.000 = 98%.

By this metric, the QC system appears to have an impressive accuracy of 98%, yet it is clear that this system is an utter failure for its primary task. In terms of statistics, its recall is 0%. **Recall**, also known as sensitivity or true positive rate, measures the proportion of actual positives that are correctly identified. It answers the question: Of all the positive cases, how many did we correctly predict? A high recall indicates that the model correctly identified most of the positive cases. **Precision**, on the other hand, measures the proportion of positive identifications that were actually correct. It answers the question: Of all the cases we predicted as positive, how many were actually positive? A high precision indicates that the model's positive predictions are trustworthy.

Let's revisit the QC system above. To ensure it catches all actual defects, we need to drive down the number of false negatives, i.e. the defective widgets the system deems of good quality. Out of the 1,000 widgets produced daily, it correctly identifies 19 of the 20 defective widgets, missing only one. However, this aggressive approach comes at a cost. The system also incorrectly flags 490 perfectly good widgets as defective, while correctly passing 490 good widgets. This results in a total of 509 widgets being flagged for inspection or rework, despite only 20 actually being defective. Such a system

demonstrates high recall, catching (TP)/(TP+FN) = 19/(19+1) = 95% of the defective widgets, but suffers from very low precision, with only about (TP)/(TP+FP) = 19/(19+490) = 3.8% of its defect predictions being accurate.

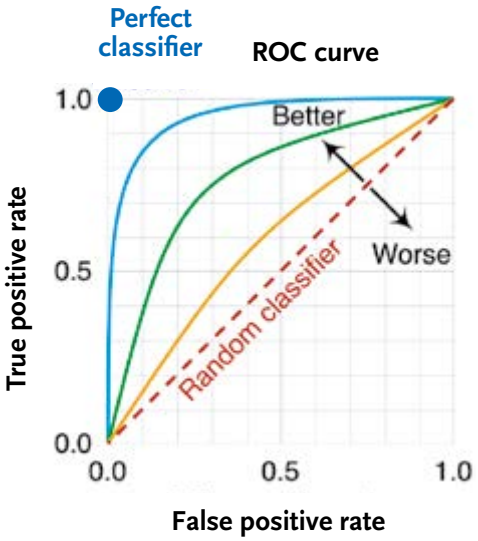
In an attempt to reduce the workload for inspectors, the QC system is adjusted to reduce its false positive rate. This new, more conservative approach results in a significant shift in the system's performance. Out of the 1,000 widgets produced daily, the system now correctly identifies only 5 of the 20 defective widgets, missing 15. However, it drastically reduces the number of false alarms, incorrectly flagging just 1 perfect widget as defective. The system correctly passes 979 good widgets without issue. As a result, only 6 widgets in total are flagged for inspection or rework, a dramatic reduction from the previous scenario. This adjustment showcases high precision, with about 83.33% of its defect predictions being accurate. However, it comes at the cost of low recall, catching only 25% of the actual defective widgets. While this approach significantly reduces the inspection workload and minimizes unnecessary rework, it increases the risk of defective products slipping through the quality control process and potentially reaching customers.

To better understand the overall performance of these two systems, we need to consider their F1 scores, which provide a balanced measure of precision and recall. The F1 score is the harmonic mean of precision and recall, providing a single metric that balances the trade-off between the two. This is particularly useful in scenarios like this, where there's an uneven class distribution (far more good widgets than defective ones) and we care about both false positives and false negatives.

For the first system, which prioritized high recall, the F1 score is approximately 7.16%. The second system, which focused on high precision, achieves a higher F1 score of about 38.46%. Both indicate suboptimal performance. But what if we had a different QC system with an F1 score of 90%? This indicates a significant improvement, but captures still quite some variation: both situations with recall of 99% and precision of 82% and with recall of 70% and precision 97% give a 90% F1 score, but would in practice have significantly different impact on quality control.

The **Receiver Operating Characteristic (ROC)** is another critical tool in understanding and evaluating the performance of classification models, especially in binary classification problems. The ROC curve plots the true-positive rate (TPR) against the false-positive rate (FPR) at various threshold settings. The ROC curve provides a comprehensive view of the trade-off between the TPR and FPR for every possible threshold. A model that perfectly distinguishes between the two classes will have an ROC curve that hugs the top left corner of the plot, indicating high sensitivity and specificity. In the chart below, that's the blue dot at 0,0/1,0. Conversely, a model that performs no better than random will have an ROC curve that's a diagonal line from the

bottom left to the top right – the red line in the chart. The yellow line performs slightly better, with a somewhat higher performance from the green line.



The key metric for an ROC curve is the area underneath it, commonly referred to as the AUC. Two classifiers may have different ROC curves but the same AUC, and in terms of overall discriminative ability, they can be considered comparable in performance. However, it's important to note that while the AUC provides a single scalar value summarizing the overall performance, the shape of the ROC curve can provide more nuanced insights. Two classifiers with the same AUC might perform differently at specific thresholds. One might be better at achieving high sensitivity, while the other might excel at specificity.

The ROC curve is particularly useful in scenarios where the classes are imbalanced or when the costs of different types of errors (false positives vs. false negatives) vary. By examining the ROC curve, one can choose a threshold that offers an acceptable balance between sensitivity and specificity for a particular application.

Steps to improve accuracy

ALTAI

B13. Could a low level of accuracy of the AI system result in critical, adversarial or damaging consequences?

A low level of accuracy of an AI system may lead to critical, adversarial, or damaging consequences in various domains. For instance, AI-driven financial trading algorithms are used to make split-second decisions on buying or selling stocks based on a myriad of factors. If such an algorithm consistently makes inaccurate predictions about stock movements, it could lead to massive financial losses for investors. An inaccurate prediction in an AI-driven

agricultural crop management systems could lead to farmers planting or harvesting at suboptimal times, leading to reduced yields.

That's not to say that a low level of accuracy is always critical or harmful to users. For example, if a movie recommendation algorithm inaccurately recommends a romantic comedy when the user typically enjoys action films, the consequence is minor. The user can simply ignore the recommendation and choose another title. In a computer game, the AI could inaccurately interpret a user's action and makes the characters or game respond in an unexpected way.

There is no generally accepted framework for this type of consequence. The AI Act mentions risks to health, safety, fundamental rights, democracy and rule of law and the environment as typical examples of situations where consequences would be critical, adversarial or damaging. However, it's essential to approach this with nuance. For instance, consider a smart trash bin that incorrectly labels recyclable waste as non-recyclable. While this does pose an environmental risk, labeling it as a "critical" consequence might be an overstatement.

When evaluating the significance of risks caused by low accuracy in AI systems, one can consider a multi-faceted approach that weighs various factors. Here's a guidance framework:

Impact on Health and Safety	Impact on Fundamental Rights	Impact on Democracy and Rule of Law	Impact on the Environment	User Dependency	Public Perception and Trust
High Significance: Systems where inaccuracies can lead to physical harm, health risks, or loss of life (e.g., medical diagnosis tools, industrial robots). **Low Significance:** Systems where inaccuracies might be inconvenient but won't harm users or their health (e.g., fitness tracking apps miscounting steps).	**High Significance:** Systems handling sensitive personal data where inaccuracies can lead to breaches, misuse, or violation of rights (e.g., surveillance systems misidentifying individuals). **Low Significance:** Systems dealing with non-sensitive data or where rights aren't directly impacted (e.g., weather prediction apps).	**High Significance:** Systems where decisions are irreversible, can influence democratic processes, or have legal implications (e.g., voting prediction algorithms, criminal sentencing tools). **Low Significance:** Systems with decisions that don't directly influence democratic or legal outcomes (e.g., public opinion survey tools).	**High Significance:** Systems where inaccuracies can lead to environmental harm or hinder conservation efforts (e.g., environmental monitoring tools misreading pollution levels). **Low Significance:** Systems where environmental implications are minimal or indirect (e.g., smart home systems slightly misadjusting room temperatures).	**High Significance:** Systems where users heavily rely on the AI's accuracy for critical tasks, impacting their rights, safety, or democratic participation (e.g., emergency response systems). **Low Significance:** Systems used for leisure or non-critical tasks that don't directly impact the aforementioned areas (e.g., music recommendation algorithms).	**High Significance:** Systems where public trust is paramount, and inaccuracies can lead to widespread skepticism, impacting democracy or public safety (e.g., public health announcement systems). **Low Significance:** Systems where public perception is less tied to accuracy and doesn't directly impact the core areas of the AI Act (e.g., art generation algorithms).

ALTAI

B14. Did you put in place measures to ensure that the data (including training data) used to develop the AI system is up-to-date, of high quality, complete and representative of the environment the system will be deployed in?

In practice, the most impactful step to ensure accuracy is to use high-quality data. The data, especially the training data, should be up-to-date, complete, and truly representative of the environment in which the AI system will operate. This is crucial because an AI system trained on outdated or unrepresentative data can produce misleading or incorrect results. Moreover, the quality of the data directly impacts the quality of the AI system's outputs. We'll delve deeper into the intricacies of data governance in Chapter 6.

ALTAI

B15. Did you put in place a series of steps to monitor, and document the AI system's accuracy?

Continuous monitoring of the AI system's performance is also essential. By regularly checking and documenting the system's accuracy, any deviations or anomalies can be detected early, allowing for timely interventions. This ongoing monitoring ensures that the system remains reliable and trustworthy over time.

ALTAI

B16. Did you consider whether the AI system's operation can invalidate the data or assumptions it was trained on, and how this might lead to adversarial effects?

It's also vital to consider the dynamic nature of data and the environment. As situations change, there's a possibility that the AI system's operations could render the data it was trained on, or its underlying assumptions, obsolete. Imagine an AI system designed for an e-commerce platform to recommend products to users based on their browsing and purchase history. The system was trained on data from the past two years, during which a particular fashion trend, let's say "bell-bottom jeans," was highly popular. Now, fast forward to the present day, where fashion trends have shifted, and "skinny jeans" have become the new rage. However, the AI system, still operating on its older training data, continues to heavily recommend bell-bottom jeans to new users, assuming that they are still in vogue. As a result, the system's recommendations might not resonate with the current preferences of the users, leading to decreased sales and user satisfaction.

Such scenarios can lead to adversarial effects, where the system starts behaving in unintended ways. Being aware of this risk and proactively checking for such shifts is crucial.

ALTAI

B17. Did you put processes in place to ensure that the level of accuracy of the AI system to be expected by end-users and/or subjects is properly communicated?

Lastly, communication is key. The end-users and subjects interacting with the AI system should be well-informed about its expected level of accuracy. Setting clear expectations helps in building trust and ensures that users are making informed decisions based on the AI system's outputs. Proper channels should be in place to communicate any updates, changes, or potential issues related to the system's accuracy to all relevant stakeholders.

Reliability, fallback plans and reproducibility

The integration of AI systems into various sectors has brought about transformative changes, offering solutions to complex problems and optimizing processes. However, with this integration comes the responsibility to ensure that these systems are both reliable and reproducible, especially in high-stakes scenarios. Reliability and reproducibility in AI systems are key to ensure consistent and predictable AI behavior.

On reliability and reproducibility

Reliability in the context of AI refers to the system's ability to consistently perform its intended function across a range of inputs and scenarios. A reliable AI system is one that users can trust to work as expected, irrespective of external variables or changing conditions. It's akin to expecting a car to start every morning; the consistent performance instills confidence.

On the other hand, **reproducibility** is about ensuring that an AI experiment or operation exhibits the same behavior when repeated under the same conditions. It's a cornerstone of scientific integrity. In AI, reproducibility means that given the same data and the same model parameters, the system will produce the same results, regardless of when or where it is run. This is crucial for validating findings, especially when AI models are used to inform critical decisions.

A common misconception is that reproducibility and reliability are interchangeable. While both are measures of consistency, they apply in different contexts. Reproducibility means producing the same result under identical conditions, in particular during testing. Reliability, by contrast, means performing well across varied real-world conditions. A system can be reliably wrong (always fails the same way), or reproducibly unstable (same inputs, erratic outcomes), so it's important to assess both.

To illustrate the sometimes subtle difference: a weather prediction model that consistently predicts rain every Tuesday is highly reproducible because it's consistent and can be replicated under the same conditions. However, the model is not reliable as its predictions are utterly disjoint from reality. At the other end of the spectrum, consider a stock trading algorithm that has been trained on a vast amount of historical data and has consistently generated profits for several months. This algorithm is reliable in the sense that it has a proven track record of making profitable trades over a specific period. However, its black-box nature and lack of documentation on training parameters, design choices et cetera make this a less reproducible AI system.

ALTAI

B18. Could the AI system cause critical, adversarial, or damaging consequences (e.g. pertaining to human safety) in case of low reliability and/or reproducibility?

B18a. Did you put in place a well-defined process to monitor if the AI system is meeting the intended goals?

B18b. Did you test whether specific contexts or conditions need to be taken into account to ensure reproducibility?

In certain domains, the stakes are exceptionally high. Consider the financial sector, where AI-driven trading algorithms manage billions of dollars. A slight inconsistency in the system's operation could lead to significant financial losses. Similarly, in urban planning, AI models predicting infrastructure wear and tear can influence decisions on maintenance and resource allocation. An unreliable prediction could result in infrastructure failures, posing risks to public safety.

Monitoring, verification and documentation

ALTAI

B19. Did you put in place verification and validation methods and documentation (e.g. logging) to evaluate and ensure different aspects of the AI system's reliability and reproducibility?

B19a. Did you clearly document and operationalise processes for the testing and verification of the reliability and reproducibility of the AI system?

To mitigate these risks, it's essential to have a well-defined process to monitor AI systems continuously. Regular checks ensure that the system is meeting its intended goals and operating within acceptable parameters. This involves not only monitoring the system's outputs but also understanding how specific contexts or conditions might affect reproducibility. For instance, an AI model trained on summer data might not be reliable in winter conditions, necessitating context-aware monitoring. Again, this goes hand-in-hand with data governance, the subject of the next chapter.

Beyond monitoring, there's a need for rigorous verification and validation methods. These methods evaluate various aspects of the AI system's reliability and reproducibility.

Proper documentation of these processes, including logging, is crucial. It provides a clear roadmap for testing and verification, ensuring that stakeholders can trust the system's operations.

Verification and validation (V&V) are critical processes in the development and deployment of AI systems, especially when these systems are used in safety-critical applications. Verification ensures that the system is built correctly according to the specified requirements. These processes by themselves are well-understood and documented in the technical literature, even when the Agile methodology is used.[15]

The role of fallback plans

ALTAI

B20. Did you define tested failsafe fallback plans to address AI system errors of whatever origin and put governance procedures in place to trigger them?

Even with rigorous monitoring and validation, errors can occur. It's essential to have tested failsafe fallback plans in place to address any AI system errors, regardless of their origin. These plans, coupled with governance procedures, ensure that when anomalies are detected, there's a clear protocol to mitigate potential damages. Not just to ensure business continuity – surprisingly, fallback plans are key to establishing trust in AI. One significant failure can erode the confidence users have in the system. By having robust fallback mechanisms, organizations can assure users that they are prepared for contingencies, thereby bolstering trust.

Fallback plans can range from sophisticated strategies like maintaining an off-site duplicate of the entire system, which continuously syncs with the primary master system, to simpler solutions. For instance, a watchdog system may monitor response times of an AI customer service system, and transfer customers to a human operator if the AI appears to be unresponsive for too long. In a smart home setting, a set of preprogrammed defaults may set in if the algorithmic environment management system makes many changes in a short time. And in a decision-support system, the AI-driven advisor can be replaced by a questionnaire or flowchart that addresses common situations.

Effective governance is the backbone of any failsafe strategy. It's not enough to have a backup plan; organizations must also define clear procedures for when and how to activate these plans.

❶ **Monitoring and Alerts:** Continuous monitoring of the AI system can detect anomalies or performance drops. Automated alerts can notify relevant teams immediately when predefined thresholds are breached.

❷ **Decision Protocols:** Clearly defined protocols should be in place to determine when to switch to the fallback system. This could be based on the severity of the malfunction, the potential impact, or a combination of factors.

❸ **Regular Drills:** Just like fire drills, organizations should conduct regular failsafe drills. This ensures that in the event of a real crisis, teams know exactly what to do, minimizing response times.

❹ **Feedback Loops:** After activating a fallback plan, there should be mechanisms to gather data on what went wrong with the primary system. This feedback can be invaluable for preventing future failures.

❺ **Stakeholder Communication:** Clear communication channels should be established to inform stakeholders about any disruptions and the activation of fallback plans. Transparency in such situations can mitigate panic and confusion.

❻ **Review and Update:** Fallback plans should not be static. They should be regularly reviewed and updated based on technological advancements,

The impact of low confidence scores

ALTAI

B21. Did you put in place a proper procedure for handling the cases where the AI system yields results with a low confidence score?

AI systems are designed to process vast amounts of data and make decisions based on patterns and information they've been trained on. However, there are instances where the system might encounter unfamiliar scenarios or data points that don't align well with its training. In such situations, the AI might produce results with low confidence scores, indicating its uncertainty regarding the decision or prediction. This uncertainty can arise from various factors, such as data anomalies, insufficient training on specific data subsets, or inherent complexities in the problem being addressed.

Confidence scores in AI systems, especially in classification tasks, typically represent the probability of a particular output or decision. These scores are derived from the underlying algorithms, with some models, like neural networks, naturally producing a probability distribution over classes. For instance, in a binary classification, a confidence score of 0,8 for a particular class means the model believes there's an 80% chance of that class being the correct one. Note that this does not mean that the model is correct 80% of the time when it gives this score. Rather, it represents the model's estimated probability for a specific instance, based on its training and the input data at hand.

Having a robust procedure to handle cases of low confidence is crucial to ensure the reliability and trustworthiness of the AI system. When the system identifies that its confidence in a particular result is below a predefined threshold, it could reroute the decision-making process to a human expert who can evaluate the situation with a more nuanced understanding. This human-in-the-loop approach (see previous chapter) ensures that critical decisions aren't made solely based on uncertain AI predictions. Alternatively, the system could be designed to seek additional data or inputs that might help in bolstering its confidence. For instance, if an AI in medical diagnostics is unsure about a scan result, it might request additional tests or scans to make a more informed decision. This proactive approach ensures that the AI system remains a reliable tool, even in uncertain situations.

A different approach is to present outputs in a tiered manner: as the confidence level decreases, the response becomes more cautious and tentative. Let's consider an AI-powered chatbot designed for customer support in an e-commerce platform. When the chatbot processes a question on the company's returns policy, the confidence in its resonse may vary, and the way it presents the answer can be adjusted accordingly:

1. **High Confidence (e.g., 90% and above):** "Electronic items purchased from our store can be returned within 30 days of purchase, provided they are in their original condition and packaging."
2. **Medium Confidence (e.g., 70% - 89%):** "I believe electronic items can typically be returned within 30 days of purchase, as long as they're in their original condition. However, I recommend checking our official return policy page or speaking with a human representative to confirm."
3. **Low Confidence (e.g., 50% - 69%):** "I'm not entirely sure, but I think electronic items might have a 30-day return window. It would be best to consult our official return policy page or connect with one of our team members for a definitive answer."
4. **Very Low Confidence (below 50%):** "I'm sorry, I'm having trouble retrieving that information right now. Would you like me to direct you to our official return policy page or connect you with a human representative?"

Continual Learning and its implications

ALTAI

B22. Is your AI system using (online) continual learning?
B22a. Did you consider potential negative consequences from the AI system learning novel or unusual methods to score well on its objective function?

Continual learning, sometimes also referred to as self-learning or lifelong learning, is a paradigm in AI where models are designed to learn continuously over time, adapting to new data without forgetting previous knowledge.16 In contrast to static, pre-trained models, continual

learning systems update themselves post-deployment, often in response to live input streams. This allows them to adjust to evolving environments or user preferences, as is common in recommendation engines, adaptive healthcare systems, or fraud detection.

While this approach makes the learning process very efficient, it comes with unique risks. Adaptive systems may discover edge-case strategies or "shortcut learning" that boost performance metrics while undermining real-world safety, fairness, or robustness strategies. For example, a recruitment algorithm might initially perform well but, over time, begin overweighting superficial proxies for success (such as writing style or formatting choices) if those features start correlating more strongly with recent positive outcomes.

The AI Act requires that any such feedback loops are duly addressed with appropriate mitigation measures. When continual learning processes lead to a substantial change in the system's behavior, a revised conformity assessment (see chapter 3) may even become necessary. The post-market monitoring mechanisms (again, chapter 3) should have particular focus on this type of unintended changes in performance over time.

A specific variant, known as online continual learning, updates the model incrementally with each new data point, often without retraining from scratch or storing historical input.[17] While powerful in dynamic contexts, this approach significantly complicates reproducibility, logging, and conformity re-assessment, and must be governed with heightened caution under post-market monitoring obligations.

To harness the benefits of continual learning while mitigating its risks, it's crucial to have robust monitoring and evaluation mechanisms in place. Regularly evaluating the model's performance, checking for drifts in objectives, and ensuring that the learning process is transparent and interpretable can go a long way in ensuring that continual learning systems remain reliable and trustworthy.

Note that continual learning is not the same as self-learning. A self-learning AI system is designed to improve its performance over time by refining its algorithms and models based on feedback loops, without the need for explicit retraining from humans. It essentially learns from its mistakes and successes, adjusting its internal parameters to optimize future outcomes. On the other hand, continual learning AI allows the AI to adapt to new data or tasks while retaining its knowledge from prior experiences. In essence, while self-learning focuses on iterative improvement within a specific domain or task, continual learning emphasizes adaptability across multiple tasks or domains over an extended period.

Key takeaways

Building robust and reliable AI systems is essential to ensuring safe, predictable, and trustworthy outcomes. This chapter explored how resilience to attacks, accuracy in decision-making, and well-defined fallback mechanisms all contribute to technical dependability. Ultimately, robustness is about the organizational ability to anticipate, detect, and respond to risk over time. This includes setting clear expectations for accuracy, evaluating the consequences of error, and ensuring systems can be verified, tested, and, when necessary, overridden.

Many of these safeguards begin with the data that AI systems are built on. In the next chapter, we turn to data governance and privacy – looking at how quality, integrity, and responsible handling of data are fundamental to building lawful and ethical AI.

6

Data Governance and Privacy in AI Systems

In a data-driven world, the importance of privacy and data governance in AI cannot be overstated. While privacy is often thought of as a 'soft' legal aspects, the European Union's General Data Protection Regulation (GDPR) has made it clear privacy and data protection is something to take seriously. The rise of generative AI has caused another legal right to raise its head: intellectual property (IP) owners are raising serious objections against the massive usage of their works in the myriad of AI systems currently being built. What does this mean for data governance?

Introduction to privacy and AI

Privacy is a fundamental human right, enshrined in national constitutions, international conventions and the EU's Charter of Fundamental Rights. Famously, privacy has been formulated as "the right to be let alone"; in legal terms, respect for one's private and family life, home and correspondence. However, privacy goes beyond just the private life. It encompasses the right to mental, physical, and moral integrity, and includes a right of self-determination regarding information about oneself. [1]

The European perspective

The European Union has long been at the forefront of championing the rights of individuals in the realm of data protection. Rooted in the belief that every individual has the right to control their personal data, the EU's 2018 General Data Protection Regulation (GDPR) has established a robust framework that set the global standard. It not only mandates stringent measures for data collection and processing but also emphasizes the importance of individual consent, transparency, and the right to be forgotten.

In the European perspective, personal life and data protection are not merely two aspects of one overarching right to privacy. Protection of personal data is a separate right enshrined in the Charter. This separate origin must be understood in the light of

By the end of this chapter, you'll be able to ...

- Navigate privacy, data protection and intellectual property laws for AI systems.
- Take steps to ensure data quality and integrity in AI systems.
- Set up and apply protocols for data acquisition and usage.

large-scale data processing by national governments (in particular in Germany) in the 1970s, which led to massive protests by citizens. The German *Bundesdatenschutzgesetz* of 1976 was the first to explicitly regulate the usage of personally-identifiable data.[2] A similar view was adopted by the OECD in 1980 with its *Guidelines on the Protection of Privacy and Transborder Flows of Personal Data*[3] and the 1981 Council of Europe's *Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data* (Convention 108), both of which can still be recognized in modern data protection legislation such as the GDPR.

This perspective is fundamentally different from the US approach to handling of "personally identifiable information" (PII). Not only is "PII" a much more limited concept than "personal data" under the GDPR,[4] the US Constitution's First Amendment puts stringent limits on any regulation of processing of such data. The California Consumer Privacy Act of 2018 (CCPA) is the first real attempt to regulate personal data, comparable in scope to the GDPR.[5]

The impact of AI

The 2010s marked the meteoric rise of big data, a phenomenon characterized by the exponential growth in the volume, variety, and velocity of data being generated. The search by businesses, governments, and institutions for tools to process has led to a renewed interest in machine learning techniques, in particular deep learning, which thrived on this data deluge. Combined with an influx of venture capital this indirectly prompted the AI revolution we are in today.[6]

The symbiotic relationship between big data and AI has reshaped the digital landscape, but not without introducing significant privacy concerns. With the capability to process and analyze vast datasets, both state and private entities gained unprecedented insights into individual behaviors, preferences, and patterns.[7] While states could justify the use of AI-driven surveillance for national security purposes, it often teetered on the brink of invasive monitoring, potentially compromising the privacy rights of citizens. On the other hand, private corporations, in their quest for profit, not only harnessed AI to dissect personal data for hyper-targeted advertising, but also significantly invested in the development of surveillance technology. AI research, especially in the realm of computer vision, has become intrinsically linked to the proliferation of mass surveillance, further blurring the ethical boundaries of data usage and privacy.[8]

The intertwining of big data with automated decision-making further compounded the many concerns over AI. Decisions that once required human judgment began to be delegated to algorithms, leading to potential biases and opaque determinations that could profoundly impact individuals' lives. In this confluence of big data and AI, the

sanctity of personal privacy stands at a crossroads, underscoring the urgent need for thoughtful and robust safeguards.

AI systems and fundamental rights

The right to privacy is a cornerstone of many legal frameworks and is deeply intertwined with human dignity. AI systems, by their very nature, process vast amounts of data, some of which can be deeply personal. This creates a fundamental challenge for AI systems, as they must be designed to respect these fundamental rights.

Challenges to fundamental rights

ALTAI

C1. Did you consider the impact of the AI system on the right to privacy, the right to physical, mental and/or moral integrity and the right to data protection?

As discussed in chapter 4, a key aspect of AI is autonomy: the ability to perform (cognitive) tasks without the need for continuous human intervention. This may contribute to a loss of agency with human operators, or more generally a challenge to their human dignity. For instance:

1. **Devaluation of human skills:** As AI systems excel in tasks once deemed complex and uniquely human, there's a risk that human skills and expertise will be undervalued or overlooked. This can lead to a diminished sense of self-worth among professionals whose roles are being replaced or augmented by AI.
2. **Erosion of personal relationships:** If human interactions are increasingly mediated or replaced by AI (e.g., caregiving robots or virtual companions), the depth and authenticity of human relationships could be compromised, leading to feelings of isolation and a devaluation of human-to-human connection.
3. **Bias and discrimination:** Autonomous AI systems that inadvertently perpetuate biases can reinforce societal inequalities. When individuals are unfairly treated by AI, not only are their rights violated, but their inherent worth as individuals is implicitly denied.
4. **Reductionist views of humanity:** There's a danger that as we come to rely on AI for more tasks, we might start to view human beings in mechanistic or reductionist terms, valuing them only for their data or as cogs in a machine, rather than as holistic, multifaceted individuals.
5. **Loss of personal narrative:** As AI systems begin to predict, suggest, and even dictate our preferences, choices, and behaviors, there's a risk that individuals may lose their sense of personal narrative and identity. Instead of life stories being shaped by personal experiences, challenges, and choices, they could increasingly be influenced by algorithmic recommendations. This could lead to a homogenization of experiences and a loss of the unique, individual stories that define our humanity and personal growth.

Identifying these challenges is very complex, as they intertwine with our deeply held values, societal norms, and the ever-evolving technological landscape. To navigate this intricate web, several approaches can be employed. Firstly, continuous stakeholder engagement can provide diverse perspectives and highlight potential areas of concern. Secondly, ethical frameworks and guidelines can offer a structured way to evaluate AI systems against established principles. Thirdly, the use of Impact Assessments, particularly the so-called Fundamental Rights Impact Assessments, can be instrumental in systematically identifying, evaluating, and mitigating potential infringements on human dignity. These assessments, along with other tools and methodologies, will be delved into in greater detail in Chapter 11.

The interplay of AI and the right to privacy

As already noted, many applications of AI can touch upon the fundamental right to privacy. A seasoned approach to AI development always places the individual's right to privacy at the forefront, ensuring that data collection, processing, and storage are all conducted with the utmost respect for this fundamental right. The interplay of AI and personal data will be the subject of the next section. But privacy-mindful AI system design is more than data protection: privacy issues can arise in many aspects of AI system design and deployment. Let's look at a few:

1. **Biometric surveillance:** The use of AI in facial recognition and other biometric tools can lead to a pervasive feeling of being constantly watched, even if no data is being stored or decisions made.
2. **Emotion recognition:** AI systems that claim to detect a person's emotional state based on their facial expressions, voice, or other cues can be invasive, potentially misinterpreting emotions and leading to unwarranted conclusions or actions.
3. **Eavesdropping devices:** Devices that are always listening for a "wake-up word" or command can inadvertently capture private conversations or sensitive information. Even when they don't, users can *perceive* this risk and feel forced to adjust their behaviour.
4. **Predictive behavior analysis:** AI systems that predict a person's future actions or preferences can be seen as invasive, making assumptions about personal choices and lifestyles.
5. **Deepfakes and image manipulation:** AI tools that can manipulate images or videos to create realistic but entirely fake content can infringe on an individual's right to their own likeness and create privacy concerns.

The advent of AI has brought about a new perspective on privacy. Unlike human interactions, where there's always a risk of personal biases and indiscretions, AI processes information neutrally. There's no "gossip" or "judgment" from a machine. It simply processes data without emotion or intent to disclose. However, this seemingly

neutral processing can inadvertently lead to heightened privacy risks for certain individuals. People with specific, private needs often find themselves at a disadvantage. For instance, pregnant women, the economically disadvantaged, or men with bladder issues might need to reveal more personal information to interact with a system that others use seamlessly. This constant revelation of private details can lead to feelings of vulnerability and discrimination, emphasizing the need for AI systems to be designed with empathy and inclusivity at their core (see also chapter 8).

Upholding physical, mental, and moral integrity

AI systems, particularly those that engage directly with individuals, wield significant influence over a person's physical, mental, and moral well-being. For instance, think about AI-driven healthcare tools offering diagnostic recommendations or social robots designed to interact with the elderly. The potential repercussions of these technologies on an individual's health and mental state can be immense.

Addressing this influence is a fundamental step in the design of an AI system. In this particular case one must always seriously consider the question of whether the AI system *should* even perform the activity at all.

1. **Misdiagnosis by AI healthcare tools:** Relying solely on AI for medical diagnoses can lead to incorrect treatments, potentially endangering lives. To mitigate this, it's crucial to implement rigorous testing and validation processes and always involve human oversight in critical healthcare decisions.
2. **Over-reliance on social robots leading to isolation:** Depending too much on AI companions (see chapter 3) can reduce human-human interactions, leading to feelings of loneliness and isolation. Systems should be designed to encourage human interaction, and usage limits should be set to prevent over-dependence.
3. **Mental stress from AI monitoring systems:** Continuous monitoring by AI can lead to anxiety and a feeling of being constantly watched. To address this, it's essential to ensure transparency in AI monitoring and provide users with the ability to turn off non-essential monitoring – with clear visual or other confirmation and safeguards that the system indeed is not monitoring anymore.
4. **Physical harm from autonomous machines:** Machines that operate autonomously can malfunction, posing direct physical risks to users or bystanders. Incorporating multiple safety redundancies and emergency shut-off mechanisms can help prevent such incidents.
5. **Moral dilemmas from AI decision-making:** AI systems making decisions in morally ambiguous situations can lead to outcomes that conflict with human values. Establishing ethical guidelines for AI behavior and involving diverse human input in moral decision-making processes can guide the system towards more ethically sound decisions.

Mechanisms for flagging privacy concerns

ALTAI

C2. Depending on the use case, did you establish mechanisms that allow flagging issues related to privacy concerning the AI system?

Despite our best efforts in carrying out impact assessments and designing with fundamental rights in mind, an AI system may still exhibit issues related to privacy or protection of personal data. Therefore, it's essential to have a flexible yet robust mechanism in place that allows stakeholders to flag potential privacy concerns. Such a mechanism should be intuitive, easily accessible, and should ensure that concerns are addressed promptly. However, it's not just about having a system in place; it's about fostering a culture of vigilance and responsiveness. Here are three practical tips to foster such a culture:

1. **Open channels of communication:** Establish clear and open channels for customers and employees to voice their concerns, ask questions, or report potential issues without fear of retribution. This could be in the form of regular town hall meetings, anonymous suggestion boxes, or dedicated forums. When employees feel that their concerns are heard and valued, they are more likely to be proactive in flagging and addressing potential privacy issues.
2. **Supplier dialogue:** Cultivating a strong and transparent relationship with the producer of an AI system is paramount. Open dialogue fosters mutual understanding and trust. This goes beyond legal measures such as a service level agreement or security review; those establish baselines but do not foster trust and cooperation. When both parties are aligned in their understanding and objectives, addressing privacy concerns becomes a joint effort. Moreover, a good relationship means that in times of unforeseen challenges or urgent issues, communication lines are already open, and both parties are more inclined to work together swiftly and efficiently.
3. **Celebrate proactiveness:** Recognize and reward employees who demonstrate a proactive approach to privacy, whether it's by identifying potential vulnerabilities, suggesting improvements, or simply being consistently diligent in their roles. By celebrating these actions, you not only encourage the individual but also set a positive example for others to follow, reinforcing the importance of a vigilant and responsive culture.

The GDPR and its impact on AI

Adopted in 2016, the GDPR is the flagship European legislation to protect personal data. As many AI systems process personal data, both during training and in use, understanding and applying the GDPR is a key aspect of AI compliance.

Applicability of the GDPR to AI systems

ALTAI

C3. Is your AI system being trained, or was it developed, by using or processing personal data (including special categories of personal data)?

The AI Act explicitly confirms that the GDPR is applicable in full to any processing of personal data by an AI system. So let's have a look at both these terms and how they relate to AI systems.

- **Personal data:** For the definition of "personal data" we need to look at the GDPR itself: "any information relating to an identified or identifiable natural person" (art. 4(1)). This is an extremely broad definition, especially considering the second part of the definition which explains that "an identifiable natural person is one who can be identified, directly or indirectly". This goes way beyond merely being able to assign a name or contact information to a data point. Any identifier, such as an identification number, location data, an online identifier or factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person are sufficient.

  A good rule of thumb is to assume that when datasets deal with humans, the data is personal data unless it can be convincingly established that the data is fully anonymized without any hope of tracing individual items to the 'data subject' from which they originated. But note: the term 'anonymized' does not merely mean removing names, identification numbers and the like. The GDPR calls this "pseudonymization", a process that can in theory be reversed as long as the identifiers still exist somewhere.

- **Special personal data:** The AI Act in several points refers to "special" or "sensitive" personal data. The GDPR marks certain categories of personal data as "special": personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation (art. 9(1) GDPR). Their extremely sensitive nature makes that the GDPR bans any and all processing of this type of data, except and to the extent explicitly permitted elsewhere.

  AI systems may well process such special personal data: many AI systems provide healthcare advice, monitoring or support, for instance. To name a different example, a companion app may deduce a user's sexual orientation through interaction. A car navigation system may record weekly trips to a church, mosque or synagogue, which indirectly reveals information on religion. Whether and how such information can be used, requires a careful analysis under GDPR guidance and case law and is outside the scope of this book.

Uniquely, the AI Act permits the use of special personal data in order to train, validate and test datasets for potential negative bias against natural persons (art. 10.5). All special measures required by the GDPR for such processing apply in full.

- **Data processing:** The GDPR applies to any 'processing' of personal data in an automated system (and to certain forms of processing in non-automated systems, but those are out of scope for this book). The term 'processing' again is defined extremely broadly: any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction. In short: any touching of data or input related to individual humans will trigger the GDPR.

- **Ground for processing:** Processing personal data is permitted only under one of the legal grounds listed in article 6 GDPR. Most AI providers rely under the ground of "legitimate interest", invoking their business interest in creating innovative and useful AI models. This ground requires a balancing test against the privacy rights of affected individuals. In its Opinion 28/2024, the European Data Protection Board has indicated this is in principle possible but with a large number of caveats, such as the public nature of the data involved, the moment an opt-out is offered and whether special personal data (such as sexual preference or ethnic background) is involved.

GDPR compliance measures for AI systems

ALTAI

C4. Did you put in place any of the following measures some of which are mandatory under the General Data Protection Regulation (GDPR), or a non-European equivalent?

When an AI system is processing personal data, the system must fully comply with the GDPR. This legal requirement exists independently of the AI Act's requirements: an AI system may fully meet the stringent requirements for high-risk AI yet violate the GDPR on some points, or vice versa. While GDPR compliance is its own specialism, let's have a look at the main requirements. The French GDPR supervisor CNIL has published AI how-to sheets that may provide further insights on GDPR-compliant AI deployment.[9]

ALTAI

C4a. Data Protection Impact Assessment (DPIA);

A Data Protection Impact Assessment (DPIA) is a systematic process designed to evaluate the potential risks associated with data processing activities, especially when introducing new technologies. Under GDPR, a DPIA is mandatory when a new form of processing is

"likely to result in a high risk" to natural persons or their fundamental rights. If the DPIA reveals that the system "would result in a high risk in the absence of measures taken to prevent them", prior permission from supervisory authorities must be sought.

To make the determination of "likely high risk", the GDPR offers a complex set of factors, supplemented by various guidances by supervisory authorities. Arguably, any AI system would satisfy these requirements, but ultimately it is up to the data controller to make the evaluation. Many data protection authorities have indicated a DPIA will generally be necessary. The AI Act itself is silent on this requirement, except for noting that a DPIA can be combined with the AI Act's instrument of a fundamental rights impact assessment or FRIA (see chapter 10).

ALTAI

C4b. Designate a Data Protection Officer (DPO) and include them at an early stage in the development, procurement or use phase of the AI system;

A data protection officer (DPO) is an independent officer that oversees data protection strategies and implementation. Their job entails training, awareness and monitoring compliance. Involvement of a DPO in a DPIA is required, and a DPO should be able to address the aforementioned GDPR-specific issues related to AI systems. In chapter 10, we will dive into the role of the AI compliance officer and their relationship to the DPO.

ALTAI

C4c. Oversight mechanisms for data processing;

The GDPR contains various requirements aimed at establishing oversight. The most general requirement is to have "appropriate technical and organisational measures to ensure and to be able to demonstrate that processing is performed in accordance with the GDPR" (art. 24 GDPR). Note that this is not just about *compliance* but also on *being able to demonstrate* compliance. Similar requirements exist in the AI Act: a risk management system, human oversight and so on. Effective oversight mechanisms are essential for monitoring and controlling data processing activities within AI systems. These mechanisms ensure that data processing remains transparent, accountable, and in line with established data protection principles.

ALTAI

C4d. Measures to achieve privacy-by-design and default;

Two specific mechanisms that seek to enforce easier GDPR compliance are called privacy by design and by default. Privacy by Design is a proactive approach that integrates data protection principles into the initial design and architecture of systems, processes, and practices, rather than adding them as an afterthought.[10] It ensures that privacy is a foundational element throughout the entire lifecycle of any project or initiative. Somewhat related, Privacy by Default ensures that the strictest privacy settings are automatically applied to

a system or service upon a user's first use, without requiring any manual adjustments by the user. It guarantees that personal data is only processed with the minimal necessary extent and duration, safeguarding user information from the outset.

In contrast, the AI Act does not explicitly call for any "compliance by design" or dictate defaults. AI providers must ensure their systems are compliant and have risk and quality management systems in place. Inherent in the design of the AI Act (product safety) is a by-design approach: compliance with the AI Act forces providers and deployers to implement clear mechanisms to ensure compliance. Moreover, the use of formal standards allows a clear establishment of the fact that an organisation is AI Act compliant.

ALTAI

C4e. Data minimisation, in particular personal data;

Data minimisation is the practice of limiting data collection and retention to what is strictly necessary for the intended purpose. In the context of AI, this means ensuring that only relevant and essential data is processed, thereby reducing the potential for misuse and enhancing data protection. The AI Act does not itself require data sets to be minimized, but does call for data to be 'relevant', have good quality and be regularly updated.

The GDPR thus may take the lead here in requiring certain older data to be removed from data sets, unless overriding interests for data quality, completeness and validation (e.g. bias prevention) can be shown. This will create immense tension with AI providers, who over a decade have experienced that quality will generally vastly improve with larger data sets, and whose systems rarely even have the ability to erase individual lines from data sets.

ALTAI

C4f. Did you implement the right to withdraw consent, the right to object and the right to be forgotten into the development of the AI system?

The GDPR emphasizes several user rights, including the right to withdraw consent, the right to object, and the right to be forgotten. Integrating these rights into the AI system's development ensures that users maintain control over their data and can exercise their rights as needed. User rights can be invoked against the deployer of an AI system, e.g. when old interactions are recorded and the user wants them removed. But these rights also extend to the underlying data sets from which AI systems are built. This usually poses extreme technical challenges for the data processing pipeline. For generative AI a unique extra issue exists: can these rights be invoked against synthetic output that contains mistakes about persons? What kind of measures should be built in to address that type of error?

Consideration of data lifecycle implications

ALTAI

C4g. Did you consider the privacy and data protection implications of data collected, generated or processed over the course of the AI system's life cycle?

Every piece of data processed by an AI system has a lifecycle, from collection to deletion. It's imperative to consider the privacy and data protection implications at each stage of this lifecycle. We will discuss this further below under "Ensuring Data Quality and Integrity". Labadie and Legner have created a reference model for data lifecycles that incorporates GDPR requirements.[11]

Non-personal data implications

ALTAI

C5. Did you consider the privacy and data protection implications of the AI system's non-personal training-data or other processed non-personal data?

When data does not relate to humans, it would not be regarded as personal data. Yet, such data may still have impact on humans. Imagine a pollution management system in a factory, that calculates the optimal moment for releasing polluting substances into the atmosphere. If the system releases pollutants at times when people are most active outdoors, it could lead to increased health issues, from respiratory problems to allergic reactions. Even in cases where no personal data is processed directly, privacy implications may still arise, for example when non-personal or synthetic data can be linked back to individuals, or when models infer or generate sensitive content.

The governance of non-personal data is regulated by the Data Act, which provides rules on access to and sharing of data, from connected devices, industrial equipment, IoT platforms and the like. Users of such devices have a right to access and port their data, obligations for data holders to share data under reasonable terms, with safeguards against abuse by dominant platforms.

General-purpose AI and data governance

General-purpose AI (GPAI) systems introduce a new layer of complexity to data governance. Unlike traditional AI tools developed for a specific task or context, GPAI models are trained on vast, uncurated datasets and made available for broad downstream use. This generality means that the risks and responsibilities associated with data sourcing, documentation, and privacy cannot be confined to a single application.

Data documentation and governance obligations

The AI Act provides separately regulatory regime for data governance in general-purpose AI (GPAI) systems, recognizing that the scale and opacity of modern training pipelines pose significant risks – not only to transparency, but to privacy, intellectual property, and fundamental rights. These systems are not narrowly tailored tools; they are trained on web-scale datasets and intended for broad reuse. As such, the data they rely on, and how that data is governed, has become a critical point of legal scrutiny.

GPAI providers must:

1. Draw up and make publicly available a sufficiently detailed summary of the training content, and
2. Include in their technical documentation detailed information on the type, provenance, and curation methodologies used for training, validation, and testing.

The technical documentation must specify:

- The nature and origin of training data (e.g. scraped web data, licensed corpora, synthetic inputs)
- The methods used to clean, filter, or deduplicate data
- The number of data points, their scope, and defining characteristics
- Any processes used to assess the suitability or unsuitability of sources
- Techniques applied to detect identifiable biases or data quality risks

These requirements aim to surface governance risks hidden within opaque datasets. Unfortunately, the imprecise wording here has yet to be worked out in detail. The European Commission struggled with drafting more specific (binding) codes of practice together with the large GPAI providers, who saw too much clarity on training data as a legal liability and a competitive threat.

Tools for structuring dataset documentation

To operationalize these obligations, several documentation frameworks have emerged from the research and AI ethics communities. Although voluntary, they align well with regulatory requirements. The table below compares the three most widely referenced tools:

Tool	Origin & Focus	Key Features
Datasheets for Datasets	Proposed by Gebru et al. (2018)[12] to bring a supplier-style standard to datasets. Focused on dataset-level ethics, design, and intended use.	■ Dataset motivation and composition ■ Collection process ■ Recommended uses ■ Limitations and ethical considerations
Data Cards	Developed at Google Research (Pushkarna et al., 2022)[13] Targeted at practitioners. Focuses on usability and interpretability.	■ High-level summary ■ Intended use and misuse ■ Data selection criteria ■ Known gaps or risks
Model Cards	Proposed by Mitchell et al. (2019)[14] Technically model-focused, but often includes dataset summary fields.	■ Model performance across groups ■ Intended domains ■ Training data overview ■ Ethical risk disclosure

While each tool serves a slightly different purpose, they converge on a shared goal: to bring structured visibility to the datasets and decisions that shape AI model behavior.

Data quality and representativeness

General-purpose AI models are only as good as the data they are trained on, but the scale and opacity of these datasets pose unique risks. The AI Act however does not set specific quality requirements for GPAI datasets, but only general transparency obligations (see previous subsection), including requirements to document "measures to detect the unsuitability of data sources" and "methods to detect identifiable biases."

The challenge, however, is that GPAI training pipelines often rely on scraped, unlabeled, or user-generated content gathered from across the internet. These sources are:

- Unstructured: Rarely organized or balanced for representativeness;
- Noisy: Contain duplicates, misinformation, outdated data;
- Biased: Reflect cultural, linguistic, and demographic over- and under-representation;
- Legally ambiguous: Including potentially copyrighted, defamatory, factually incorrect and privacy-sensitive material.

This introduces governance risks at two levels:

1. Internal model behavior, including toxic output, discriminatory patterns, or hallucinations.
2. Downstream accountability, especially when GPAI is reused in high-risk domains like education, employment, or access to public services.

A complication is the given that GPAI is designed for downstream incorporation. Thus, these models are not designed for a single intended population, but are nonetheless used by diverse users, often in sensitive contexts. As such, providers face an emerging duty of anticipatory governance: they must consider not just what their model is intended to do, but what it may plausibly be used for, and whether its training data enable or endanger those uses.

A downstream integrator may not have direct access to the training datasets used to build a GPAI model. Yet, it may quickly become necessary to evaluate and document data quality when working with a GPAI system. Here are common strategies to do just that:

1. Acquire and review dataset summary and technical documentation. This allows you to verify alignment between the stated dataset characteristics and your intended deployment context, and flag any major mismatches (e.g., model trained only on English data but deployed in multilingual environments).
2. Conduct targeted prompt testing ("red teaming"). This includes crafting prompts that simulate edge cases, minority perspectives, or sensitive traits, and comparing outputs for fairness, coherence, and factuality. exposure.
3. Run subgroup bias diagnostics using audit tools. Tools like Fairlearn, Aequitas, or Facets help identify disparities in model outputs across sensitive attributes (e.g., gender, age, background) and uncover unintended bias patterns relevant to your deployment context.[15]
4. Check for data lineage in fine-tuned models. If you or a vendor adapt a GPAI model using additional data, document the origin, characteristics, and scope of that dataset. Use tools like Datasheets for Datasets, Data Cards, or MLflow[16] to track provenance, versioning, and curation steps. This supports accountability, reproducibility, and compliance with AI Act documentation duties.
5. Establish output logging and version control. Maintain structured logs of model prompts, outputs, and system versions, especially if your integration supports user interaction or dynamic updates. Use tools like Weights & Biases, Neptune.ai, or standard MLOps stacks to track model behavior over time.[17] This enables traceability, supports incident response, and fulfills AI Act requirements for logging and post-market monitoring.

When working on high-risk AI, integrate these findings into your conformity assessment dossier, ideally using structured documentation formats like the Use Case Cards (see Chapter 3).

GDPR tensions: hallucinated and inferred personal data

As general-purpose AI (GPAI) models are often trained without explicit intent to process personal data, their sheer scale and generality make privacy and personal data breaches an emergent property of their design. First, there's the general concern that operators of these models include large swathes of the world-wide web, including vast amounts of personal data, into training data without as much as a notification to the persons involved. But there's more.

Studies have shown that large language models trained on web-scale data can reproduce strings of personal data, including email addresses, financial records, or names from leaked datasets. This is known as direct memorization or regurgitation of data seen during training. More complicated is the phenomenon of contextual hallucination, where the model generating plausible but false personal data. In a famous incident, Norwegian journalist Arve Hjalmar Holmen filed a defamation lawsuit against OpenAI after its GPAI system ChatGPT indicated he had murdered both his children (Homen reports on serious crimes, including against children).

Both come with the concern that affected individuals may be unaware their data was used, and thus be unable to correct or erase it, undermining GDPR rights to information, correction (Art. 16), and erasure (Art. 17). But even when individuals are aware, the way in which they could exercise those rights is fundamentally unclear. GPAI models are not databases in the classic sense, where information is looked up and processed to fit the desired output format. Rather, they are probabilistic systems that generate each output on the fly by predicting the next most likely token, based on patterns learned from training data. This makes it nearly impossible to trace a specific output to a specific input, let alone locate, modify, or erase individual data points post hoc.[18]

Modern GPAI models do extend this basic architecture with remarkable advances: they integrate multimodal inputs (text, audio, image), provide multi-step reasoning, access long contexts and support interactive memory or external tool use. While often presented as a significant advance over mere "next token predictors", these enhancements do not introduce persistent, structured access to personal data in a way that enables GDPR rights to be exercised.

Downstream accountability for GPAI integration

As GPAI is increasingly embedded in high-risk domains such as recruitment, education, public services, and healthcare, the question of who is responsible for governance lapses becomes more complex. The AI Act assigns primary obligations for GPAI documentation and transparency to their providers, but the risks of inappropriate output, biased decision-making, or privacy infringement often materialize downstream, at the point of use. This makes downstream integrators and deployers co-responsible for evaluating, adapting, and documenting how GPAI systems function within their specific use case.

A downstream integrator who finetunes a GPAI model for a high-risk purpose by law becomes the provider of the resulting high-risk AI system (see chapter 2). This holds true regardless of whether the integration occurs through:

- Local deployment (e.g. downloading the GPAI model and fine-tuning it on internal data); or
- Remote access via API or similar mechanism.

What matters is not the technical method of access, but the shift in intended purpose and control. Thus, a well-designed prompt that specifically applies a GPAI model to a high-risk purpose may be enough to qualify the organisation using it as the provider of a high-risk AI system, even when no data is added.

Use of GPAI is often informal and organic, and therefore hard to track down from a compliance perspective. Of particular concern is data quality: as noted above, GPAI providers are not held to the same high bar for data quality and representativeness as providers developing high-risk systems. Yet once a GPAI model is integrated into a high-risk application, the downstream actor inherits full responsibility for ensuring that the resulting system is trained and tested on relevant, representative, and unbiased data.

This presents a paradox. The downstream integrator is legally required to ensure data quality in a model whose original training corpus is inaccessible, undocumented, or inherently biased. The task is, in many cases, unrealistic without full visibility into upstream datasets. Mitigation strategies exist, but offer only partial protection. They can detect symptoms of underlying data issues, but they cannot cure the structural opacity of GPAI training pipelines. In practice, this means that downstream organizations integrating GPAI into high-risk systems must treat such deployments as inherently risky and document their governance decisions accordingly.

Intellectual Property (IP) and data governance

As AI systems increasingly rely on large-scale data for training and finetuning, questions of intellectual property (IP), licensing, and dataset provenance have moved from peripheral technical concerns to core data governance and compliance challenges. Unlike traditional software components, training data is rarely structured, permissioned, or auditable in the same way. Risks are twofold: first, that training or deploying a model involves unauthorized use of protected works (e.g. copyrighted texts, images, music, or code); and second, that the resulting outputs may infringe on exclusive rights or trigger reputational liability.

Copyright in the data-driven AI era

Copyright law protects creative works such as books, images, videos, software code, and even blog posts or product descriptions. Traditionally, infringement involved unauthorized reproduction or distribution of a work. While machine learning – the technology underlying modern AI – does require a vast amount of works, its type of use is utterly incomparable to the past.

Modern AI models are trained on datasets composed of billions of text snippets, images, or recordings, many drawn from web scraping or user-generated platforms. In most cases, these sources include protected works that are copied in part or whole into the training pipeline. Even if models do not store or output full works, they may internalize patterns, structures, or stylistic features that later reappear in generated content. This raises the possibility of infringement via direct reproduction.

There is also growing attention to stylistic mimicry, where models generate content in the recognizable voice, visual style, or structure of a specific artist or author. This raises novel legal questions: Is a model that mimics a creator's style infringing on moral rights or violating unfair competition laws? Does reproduction require copying of the work, or is indirect emulation enough to trigger legal protection?

Especially when using third-party generative GPAI models or systems, an unsettled question is who is responsible for any claims of copyright infringement. For an average user, determining whether AI output infringes is an impossible question. But the same holds for GPAI providers: they cannot oversee what their models output, or even connect outputs to particular inputs. That said, the base assumption in the law is that whoever reproduces or publishes is legally liable for infringements. This puts the legal risk with the user. Fortunately for them, many large GPAI providers include IP indemnification in their service agreements.

The EU framework: the Text and Data Mining (TDM) regime

To address the legal ambiguity surrounding the use of protected content in data-intensive applications, the European Union introduced a Text and Data Mining (TDM) framework in the 2019 Copyright in the Digital Single Market (CDSM) Directive.[19] Under the directive, two regimes exist:

- TDM for scientific research. Research organizations and cultural heritage institutions may perform TDM on any content they have lawful access to, without needing explicit permission from rights holders or the paying of royalties.
- TDM for general use. TDM is permitted for any purpose, including commercial use and without royalties, provided that the content has been lawfully accessed. However, this right is conditional: rights holders may opt out of this use in advance by means of machine-readable signals.

Most datasets for AI training are built on the general use regime, which has prompted large content publishers to find ways to signal their opting-out. Foreseeing that the opt-out indication would need to be recognized by automated processes, the Commission added a requirement that this indication be machine-readable. But since its introduction, debate has raged over what this means and how to implement it. Without technical standards (that do not exist), algorithms cannot reliably interpret opt-out indications. Yet, without any means to indicate an opt-out, the conditional nature of the right (and thus the rights of creative workers) is fundamentally undermined.[20]

In the EU context, work is underway to establish binding codes of conducts for providers of GPAI models, which will include specific rules on disclosing data origin and manner in which opt-out mechanisms are respected. No specific standards have been created yet, however, and current open-source or publicly available datasets (e.g. LAION, Common Crawl) do not document TDM opt-out compliance.

A related concern is that publicly available datasets are not always created from legal sources. Famously, in a 2025 US copyright lawsuit, it was found that Meta relied on e-books from the book piracy site Library Genesis (LibGen).

Practical steps for IP governance

Managing intellectual property (IP) risk in AI systems is no longer just a legal exercise, but a data governance obligation. Organizations that develop, adapt, or deploy AI systems using externally sourced data must now treat dataset licensing, copyright status, and opt-out compliance as auditable parts of the development lifecycle.

The following governance steps offer a practical baseline for organizations seeking to align their AI practices with both legal and ethical expectations in 2025:

❶ **Identify Content Types and Protected Works.** Start by classifying the types of content used in training, validation, or testing:

- Literary or journalistic text (subject to copyright and possibly licensing)
- Images and illustrations (often rights-managed)
- Source code (frequently dual-licensed or copyrighted)
- Audio or video content (subject to both performance and fixation rights)

❷ **Assess Source Legality and Access Rights.** Ensure that the data used for training or adaptation was lawfully accessed and retained. This includes:

- Reviewing terms of service or licenses from scraped or downloaded content
- Confirming lawful access in the case of purchased datasets or APIs
- Flagging any sources that indicated opt-outs (e.g., via robots.txt, noai, or noindex headers) and consider their legal impact

❸ **Document Licensing Status and Restrictions.** Create a record for each dataset or content type stating:

- Whether it is open, licensed, or restricted (when no information is available, assume "unsafe")
- Any obligations for attribution, non-commercial use, or share-alike
- Whether rights holders opted out of data mining
- Whether indemnity or usage terms apply (especially for vendor-provided data)

❹ **Reflect IP Risk in Technical Documentation.** For high-risk AI systems or general-purpose models, ensure that technical documentation includes:

- The provenance of training data, including known licenses or restrictions
- A list of datasets used or dataset classes (e.g., news, code, academic text)
- A statement of compliance with copyright and TDM obligations
- Any known legal uncertainties (e.g., unverified public domain status)

❺ **Mitigate and Monitor Risk.** Where IP risk is identified, organizations should:

- Filter or exclude flagged datasets during fine-tuning
- Use dataset registries or opt-out filters
- License commercial datasets from providers that offer indemnification
- Maintain version-controlled logs of dataset composition, licensing assumptions, and updates

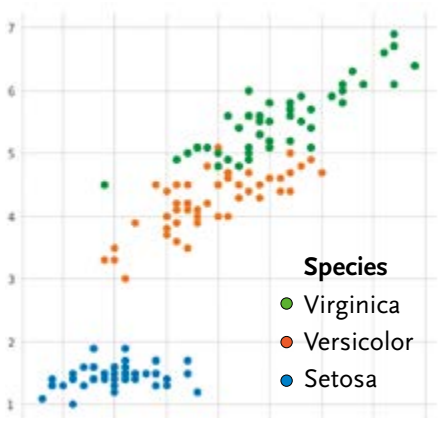
Ensuring data quality and integrity

High-quality data is a foundational requirement for the performance, safety, and fairness of AI systems. While ensuring data quality has always been a concern in machine learning, the context has shifted. Organizations are now expected to treat data sourcing, annotation, and validation not just as technical steps, but as documented governance processes. This is especially important in environments where datasets are compiled from external sources, include personal or sensitive information, or are adapted from GPAI models.

Data sets and data processing

A dataset or data set is the general term for any collection of data from which an AI model is created. A simple example is reproduced in the following table: This table provides a snapshot of the well-known 1936 Iris dataset,[21] showcasing 10 **data items** from all three species of the iris flower: Setosa, Virginica, and Versicolor. Each row represents a unique iris flower, and the columns detail the **features** of the dataset. In machine learning, the species will be the **target feature**, the label to be predicted given the other features. This is done by analyzing data: which characteristics or ranges correlate with which target characteristics? The figure below the table shows what such a correlation might look like.

Sepal Length (cm)	Sepal Width (cm)	Petal Length (cm)	Petal Width (cm)	Species
5.1	3.5	1.4	0.2	Setosa
4.9	3.0	1.4	0.2	Setosa
6.7	3.0	5.2	2.3	Virginica
6.3	2.5	5.0	1.9	Virginica
6.5	2.8	4.6	1.5	Versicolor
5.7	2.8	4.5	1.3	Versicolor
5.8	2.7	5.1	1.9	Virginica
6.0	2.7	5.1	1.6	Versicolor
5.4	3.4	1.7	0.2	Setosa
5.6	2.9	3.6	1.3	Versicolor



When working with data sets in AI systems, three key terms will make frequent appearances:

1. **Training Data:** The training data is that set of data on which the machine learning algorithms perform analysis, deriving key insights and criteria with which new outputs can be generated. It's imperative that this data is not only vast but also diverse and representative of the real-world scenarios the model will encounter. A model

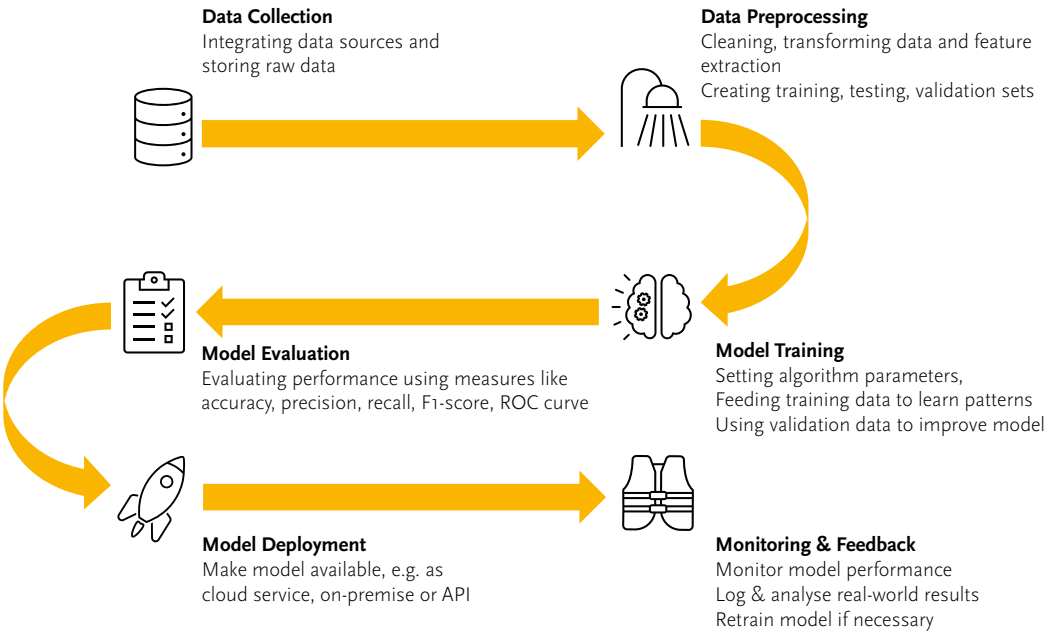
trained on a narrow or biased dataset will inevitably produce skewed results. Ensuring a broad spectrum of patterns in the training data is crucial for models to learn and adapt effectively, capturing the nuances and complexities of varied inputs.

❷ **Validation Data:** Once a model is trained, validation data steps in to refine it further. This dataset plays a crucial role in fine-tuning the model, helping it generalize well to scenarios it hasn't been explicitly trained on. Without a robust validation set, there's a risk that the model might overfit to its training data, becoming too specialized and failing to perform well in real-world applications.

❸ **Testing Data:** The final litmus test for any AI model is its performance on testing data. This dataset provides an unbiased evaluation of the model's readiness for deployment, assessing its accuracy, reliability, and overall performance. It's the checkpoint that ensures the model not only has learned well but is also prepared to deliver consistent results in diverse operational environments. Both testing and validation data should not overlap with the training data, as this will significantly affect the AI model's quality.

On data processing pipelines

In today's data-driven world, the concept of a data processing pipeline has become prevalent. The phrase refers to a set of processes and tools used to move data from one system to another, typically involving stages of data collection, processing, storage, and analysis. Think of it as a conveyor belt for data, where raw information enters at one end and emerges as actionable insights or processed data at the other. Just as an



industrial pipeline carries fluids through a series of processes, a data pipeline transports data through various stages of transformation and validation.

A data pipeline for a typical machine learning (AI) system involves several stages, each of which processes and transforms the data to make it suitable for model training and deployment, as illustrated in the figure overleaf.

Towards high quality datasets

The data that is fed into AI systems plays a pivotal role in determining their efficacy and reliability. Data quality is an upper bound to the system's quality. The need for growth has however generally won in the past years: a system improves as much, if not more, with a large addition of low- or medium quality data than with the addition of a small highly curated set. The internet offered vast amounts of low-quality data, thus seemingly obviating the need for high-quality datasets. The tide is turning however.

Still, there is much work to do. While, as noted above, good data processing pipelines are available to streamline the process of creating an AI system, the curating of the data that goes into such a system is still very much bespoke manual labor. Often, this work is outsourced to unskilled freelance operators, e.g. through the Amazon gig platform *Mechanical Turk* where workers are available to label data for a fraction of a cent per data point. Needless to say, the quality and consistence of their work is often heavily criticized.[22] Similarly, creating data sets by harvesting large quantities of data from the internet – a key practice for foundation models – is also prone to risks.[23] Curated public datasets have long been hailed as the solution, but independent research often reveal significant biases or other errors.[24]

Given the impact of AI based on such low-quality data, it is understandable that the AI Act heavily leans on the quality of data sets. However, what exactly *is* 'quality' in terms of data? The Act refers to "high quality" data, and requires "appropriate data governance and management practices" to ensure that "appropriate statistical properties" are present, in particular to avoid negative bias or discrimination towards certain groups. Nowhere are any of these terms defined, leaving it up to implementers to work out quality criteria and processes.[25]

In the previous chapter, we discussed statistical measures (accuracy, F1 and ROC). However, a dataset with the highest accuracy or F1 score may still be of low quality. Typical causes include:

1. **Overfitting:** Overfitting occurs when a statistical model or machine learning algorithm is focused too much on the precise properties of the training data, rather than all possible data. Any anomalies and outliers in the training data are treated as ordinary

occurrences instead of discarded as the noise they properly are. As a result, while it may have high accuracy on the training data, it performs poorly on new, unseen data.

❷ **Underfitting:** Underfitting, on the other hand, is the opposite problem. It arises when the model is too simple to capture the underlying structure of the data. In this case, the model fails to capture important patterns in the training data, leading to a poor fit.

❸ **Class Imbalance:** This occurs when the classes in a dataset are not represented equally. For instance, in a binary classification task, if 95% of the data belongs to Class A and only 5% belongs to Class B, a model might achieve high accuracy by simply predicting Class A all the time. However, this would not be a good representation of its ability to correctly classify instances of Class B.

❹ **Data Leakage:** This happens when information from the test set inadvertently gets used during the training of the model. It can lead to overly optimistic performance metrics, as the model has, in essence, already seen the test data during training.

The work of Budach et al. provides a complete and well-tested set of factors allowing an objective measure of quality for data:[26]

❶ **Consistent representation:** A dataset is consistent in its representation if no feature has two or more unique values that are semantically equivalent. For example, in a column listing countries, France should not be also represented as FR, La France or French Republic.

❷ **Completeness:** A dataset is complete if no items in the set have missing values. For instance, a temperature sensor that had a failure between 7 and 8 o'clock would have missing values for that time period. The completeness of the dataset is an explicit AI Act requirement.

❸ **Feature accuracy:** A feature is an element in a dataset, e.g. city and number of inhabitants. Real-world data tends to have errors in such features, which directly affects the quality of predictions. The AI Act requires the minimization of such errors.

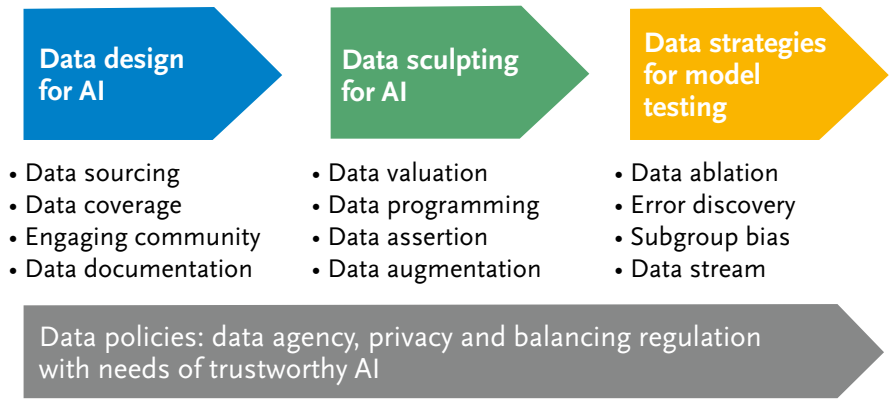
❹ **Target accuracy:** The target is that feature for which the AI system makes predictions or other output. Target accuracy thus is the feature accuracy of this particular feature.

❺ **Uniqueness:** Often, large datasets contain duplicative data, which does not improve learning and may even introduce errors (e.g. if one of the duplicates is slightly different). While exact duplicates are easy to remove, the similar-but-not-identical category is much harder to identify.

❻ **Target class balance:** In determinative AI, the target feature is a member of a class, e.g. the class 'animals' with labels such as 'cat', 'dog' or 'capibara' or the class 'approved' with labels 'yes' and 'no'. Many ML systems perform best if the dataset has an approximately equal number of labels per class, and may make mistakes if classes are highly imbalanced. For instance, if 98% of labels for 'approved' are 'yes', the system may simply always output 'yes' without further analysis and report a 98% accuracy in predicting approvals.

The quality dimensions with the largest impact are completeness, feature accuracy and target accuracy. The dimensions uniqueness and target class balance show little impact, and consistent representation has impact as soon as the new representations outweigh the old one.

Liang et al propose a new approach to creating high-quality datasets. Rather than the traditional model-centric approach, where the data set is treated as a given and effort is put mainly in optimizing the system's performance, a data-centric approach should be taken where the data pipeline is continually used to improve the data.[27] As the figure below illustrates, the process has three main steps: data design, data sculpting and model testing. The considerations mentioned in the paper are a valuable recommendation for any organization deploying a data pipeline.



The data-centric approach for high-quality data pipelines (source: Liang et al. 2022)

Confronting and addressing data biases

One of the most pressing challenges in AI today is the presence of biases in data, which can lead to skewed, unfair, and even harmful outcomes. Addressing these biases is not just a technical necessity but an ethical and legal imperative: the AI Act identifies bias as a key risk and risk management therefore must put explicit attention towards combating any potential bias. Bias in an AI system can have many causes, which we will address more generally in chapter 8. In the context of data governance we can identify a few specific sources of bias that are worthy of attention:

❶ **Sample Bias:** This occurs when the input data does not accurately represent the situation that is being modeled. A traffic prediction system trained primarily on urban traffic patterns might not accurately predict traffic in rural areas.

❷ **Association Bias:** This type of bias arises when the system incorrectly links unrelated aspects together. An advertising algorithm might associate buying sports equipment with a specific gender, leading to skewed product recommendations.

❸ **Incompleteness Bias:** This happens when the input data lacks certain crucial information. A property valuation model might not have data on recent infrastructure developments in an area, leading to undervalued property predictions.

❹ **Precision Bias:** This bias emerges when an AI's statement or prediction is mistakenly viewed as objective or of significant importance, in particular because its manner of presentation appeared very precise and accurate. A weather prediction tool might forecast a 90,6235% chance of rain, leading an event planner to cancel an outdoor event, only to find the day remains sunny.

❺ **Prejudice Bias:** This occurs when measurements or inputs are conducted or collected in a biased manner. A hiring tool might favor candidates from certain universities based on historical data, overlooking potentially qualified candidates from lesser-known institutions.

Such biases in datasets can arise from a myriad of sources. Historical prejudices, for instance, can leave lasting imprints on data. If an AI model is trained on historical data that reflects past societal biases, it can inadvertently perpetuate those biases. Skewed data collection methods, where certain groups are overrepresented or underrepresented, can also introduce biases. Additionally, unrepresentative sampling, where the data doesn't accurately reflect the broader population, can lead to models that are biased towards specific subgroups.

Combatting data biases requires a multi-faceted approach. Fairness-enhancing interventions can be employed to adjust models and ensure they make fair decisions across different groups. Adversarial testing, where models are deliberately challenged with data designed to expose biases, can help in identifying blind spots. Furthermore, sourcing data from diverse and representative sources can reduce the chances of biases creeping in. Continuous monitoring and feedback loops, where the outputs of AI models are regularly checked for biases and the models are adjusted accordingly, can also play a pivotal role in ensuring fairness.

Technical measures for data security

Given the above, the imperative to ensure data security and privacy becomes more and more pronounced. While foundational security practices – like robust password protocols, timely software patches, and firewalls – are essential, they do not fully address the unique challenges of data privacy in machine learning contexts. The process of collecting, preprocessing, training, and deploying models introduces multiple unique points of vulnerability.

Adherence to data management standards

ALTAI

C6. Did you align the AI system with relevant standards or widely adopted protocols for (daily) data management and governance?

Aligning an AI system with relevant standards or widely adopted protocols for data management and governance is paramount to ensuring the system's integrity, reliability, and compliance with best practices. One of the most recognized standards in this domain is the ISO/IEC 27001, which pertains to information security management. This standard provides a systematic approach to managing sensitive company information and ensures that robust security measures are in place to protect data from breaches and unauthorized access.

In addition to adhering to such standards, it's essential to establish a daily protocol for data management within the AI system. A typical protocol might involve routine data audits to identify and rectify any inconsistencies or errors, regular backups to prevent data loss, and periodic reviews of access controls to ensure that only authorized personnel can access sensitive data. This protocol not only ensures the smooth operation of the AI system but also reinforces trust among stakeholders by demonstrating a commitment to data protection and governance best practices.

In the research field of machine learning, many data management protocols have been developed from the perspective of ethical and reproducible experimentation.[28] Their lessons can easily find applications in today's AI environments.

Data processing techniques

Securing data during processing and ensuring data privacy of the persons affected by that data is not just a desire from lawmakers, it's also a technical challenge. Over the years, researchers have created many advanced technologies to foster better security while being able to create AI systems. Let's look at a few:

❶ **Differential privacy** is a mathematical framework that ensures the results derived from a dataset do not reveal specific information about any individual within that dataset. By introducing calibrated noise to the data or the output of a query, it guarantees that the presence or absence of a single record doesn't significantly affect the outcome. This is particularly crucial when training machine learning models on sensitive datasets, ensuring that the model's predictions don't inadvertently leak individual data points.

❷ **Homomorphic encryption** is a groundbreaking cryptographic technique that allows computations on encrypted data without requiring decryption first. In the context of machine learning, this means that models can be trained and make predictions on encrypted data, ensuring data privacy throughout the entire processing pipeline. The

resultant encrypted output can then be decrypted by the data owner, ensuring that sensitive information remains concealed from potential adversaries, including the model operators.

❸ **Federated learning** is a decentralized approach to training machine learning models. Instead of centralizing data from various sources into one location, the model is trained at the data source itself, be it a mobile device or a local server. Only model updates or gradients are shared and aggregated centrally, ensuring raw data remains at its source, significantly reducing the risk of data breaches or unauthorized access.

❹ **Secure Multi-Party Computation** is a cryptographic technique that allows multiple parties to collaboratively compute a function over their inputs while keeping those inputs private. In machine learning, this can be employed to train a model on combined data from multiple sources without any party revealing their individual data. The data remains partitioned, and intermediate computations are encrypted, ensuring data privacy is maintained throughout the collaborative process.

❺ **Data masking** involves obscuring specific data within a database, rendering it inaccessible for unauthorized users. It ensures that sensitive data remains confidential and is especially useful in development and testing environments. Tokenization, on the other hand, replaces sensitive data with non-sensitive substitutes or tokens. These tokens can then be processed without exposing the underlying data, ensuring that machine learning operations, especially in cloud environments, don't compromise data integrity or privacy.

Data storage measures

Storing data isn't just about finding a place for it; it's about ensuring that this data, whether at rest or in transit, remains inaccessible to unauthorized entities. In this section, we will explore advanced storage solutions that not only house data but also fortify it against potential breaches and unauthorized access.

❶ In the realm of data security, **tokenization** stands out as a robust method to protect sensitive information. It involves replacing sensitive data elements with non-sensitive equivalents, termed as "tokens." These tokens retain essential data characteristics without disclosing the underlying data value, ensuring that even if a breach occurs, the exposed tokens have no exploitable meaning or value.

❷ **Data masking** is another pivotal technique that aims to protect the original data. It works by concealing the actual data with altered content, yet the structure remains similar to the original. This ensures that while the data can still be used for testing and development purposes, any unauthorized access will not reveal the true sensitive information.

❸ As data breaches become more sophisticated, the need for advanced protection mechanisms has never been higher. **Encrypted databases** rise to this challenge by employing encryption techniques to safeguard data when it's at rest. By encrypting the actual data values in a database, unauthorized access will only yield indecipherable content, ensuring data remains confidential and secure.

Data access control

Securing data doesn't end once it's stored. Equally vital is the manner in which this data is accessed and utilized. Access control mechanisms serve as the gatekeepers, ensuring that data is only available to those with the right permissions, thereby preventing misuse or unauthorized access. Three organizational measures are important:

❶ **Role-Based Access Control (RBAC):** In large organizations, where myriad users require data access, RBAC plays a crucial role. It operates on the principle that not everyone needs access to all data. By assigning roles within the organization, RBAC ensures that individuals can only access data pertinent to their specific role, thereby minimizing the risk of unauthorized data manipulation or exposure.

❷ **Attribute-Based Access Control (ABAC):** While RBAC focuses on roles, ABAC takes data access control a notch higher. It defines access levels based on a combination of attributes, such as the user profile, device used, and even the time of access. This granularity ensures a more dynamic and context-aware access control, adapting to various scenarios and requirements.

❸ **Audit Trails:** Transparency and accountability are pillars of robust data security. Audit trails provide this by maintaining comprehensive logs of all data access and modifications. This is well understood as a general security measure, but becomes more prevalent in the context of AI systems: the AI Act requires logging (audit trails) of *outputs* of the AI system, allowing reconstruction of potential mistakes and ensuring a paper trail for ensuring and demonstrating compliance with the Act.

Key takeaways

Effective data governance is no longer optional in the development and deployment of AI systems. Data relevance, representativeness, and quality have become compliance requirements. This includes not only the lawful handling of personal data, but also the responsible use of copyrighted content, licensing compliance, and traceability throughout the dataset lifecycle.

Yet ensuring data quality is only part of the equation. Without meaningful transparency into how datasets are built, governed, and audited, data governance efforts cannot be verified, contested, or improved. As we turn to the next chapter, we examine the principle of transparency not as a vague ethical aspiration, but as a legal and operational requirement.

7

Emphasizing Transparency in AI Operations

In this chapter, we explore the foundational importance of transparency in building trust and the challenges surrounding its definition. The chapter examines traceability as a means of ensuring accountability, including quality of input and output and the required practices under the AI Act. We further look at the significance of explainability in AI decisions, and the ethical considerations of automated decision-making. The latter subject of course requires an examination of the GDPR and its relationship to the AI Act. We also discuss the pivotal role of communication in bridging the gap between AI systems and users, emphasizing the need for clarity and ethical interactions.

Introduction to transparency in AI

As is widely recognized, transparency is crucial for building and maintaining users' trust in AI systems.[1] In fact, the concept is the single most common principle in the vast number of ethical guidelines addressing AI on a global level.[2] Yet, there is no agreement on the actual meaning of the term. Some use 'transparency' as the opposite of the well-known "black box" AI, others see transparency as a documentation requirement or refer to the need for traceability in evaluation and decision-making.

The growing need for transparency

In the initial stages of AI development during the mid-20th century, AI systems were relatively simple. Early AI systems, such as rule-based expert systems, were designed with clear, predefined rules. These systems made decisions based on a set of explicit guidelines, making their reasoning processes transparent and interpretable. However, this changed dramatically after the rise of deep learning models, in particular neural networks. Being based on statistical models, they are great in making accurate predictions but do not have any meaningful underlying guidelines or rules to justify them. "The data says so" is an apt summary of their working.

As AI systems began to play crucial roles in sectors like healthcare, finance, and criminal justice, this so-called *black box* issue became more than just a technical challenge.[3]

By the end of this chapter, you'll be able to ...

- Explain the concept of transparency as it applies to AI systems.
- Apply best practices to establish traceability, explainability and communication.
- Address automated decision-making and work with legal limitations.

It raised ethical and legal concerns. The inability to understand why an AI made a particular decision became problematic, especially when these decisions had real-world consequences. Hence the urge for transparency in AI systems.

The "what" and the "how"

Generally speaking, transparency can take two forms: transparency on the *outcome* and transparency on the *process* of getting at that outcome.[4] The first form relates to the clarity and interpretability of the results or decisions produced by an AI system. Stakeholders, especially end-users, often need to understand the "what" behind an AI's decision. For instance, if an AI system denies a loan application, the applicant would want to know the reason for this decision. The concept of "explainable AI" (to be discussed below) relates to this aspect of transparency.

Transparency on the process refers to the clarity in understanding the mechanisms, algorithms, and data that the AI system uses to arrive at its decisions. Knowing the "how" is crucial for developers, regulators, and other stakeholders. For instance, understanding the process can help in identifying biases in the system, ensuring fairness, and making necessary adjustments. This is where documentation becomes key. Registering the entire AI development lifecycle, from data collection to model training and validation, can enhance process transparency. Additionally, using interpretable models or model-agnostic explanation techniques can shed light on the inner workings of complex models.

Transparency in both forms is a necessary prerequisite for accountability (see chapter 10). Accountability builds on the "what" and the "how" by providing the underlying "why": the ethical and legal justification for having the system operate in this way. And a black box cannot be justified.

Three aspects of transparency

Transparency can be broken down further into three aspects:

❶ **Traceability:** Traceability refers to the ability to track the decision-making process of an AI system. This includes understanding the data sets used, the algorithms applied, and the various processes that culminate in the AI system's final decision. By ensuring traceability, we can document and understand the journey of an AI decision, from the initial data input to the final output. This is especially vital when errors or unexpected outcomes arise. Being able to trace back to the root cause allows for corrective measures, continuous improvement, and accountability.

❷ **Explainability:** Explainability is about making the AI's decision-making process understandable to humans. It's not enough for an AI system to make a decision; it

must also be able to explain its decision in a manner that is clear and comprehensible, especially when its decisions have significant impacts on individuals or society. This can be challenging, especially with complex models like deep neural networks, often termed as "black boxes" due to their opaque nature. However, the goal is to strike a balance between the accuracy of a model and its explainability, ensuring that users, regulators, and stakeholders can understand and trust the AI's decisions.

❸ **Communication:** Communication is about being forthright about the AI system's capabilities and, equally importantly, its limitations. Every AI system, no matter how advanced, has its strengths and weaknesses. Communicating these openly ensures that users are aware of what the system can and cannot do. This includes informing users when they are interacting with an AI (as opposed to a human) and providing clear instructions and disclaimers about the system's use. Open communication builds informed trust, where users are not just relying on the AI blindly but are aware of its scope and potential pitfalls.

In the following sections, we will delve deeper into each of these elements, addressing specific questions and providing actionable insights to ensure the transparent and trustworthy deployment of AI systems.

Traceability: Ensuring accountability in AI systems

Traceability is key to the principle of accountability. This principle, further discussed in chapter 10, means that every decision made by an AI system can be attributed to a specific process or action within the system. By ensuring traceability, we are essentially creating a documented pathway that can be followed to understand how a particular AI decision was reached. This not only bolsters confidence in the system but also ensures that when things go awry, there's a clear trail to follow, pinpointing where and why a mistake occurred.

Traceable lifecycle

ALTAI

D1. Did you put in place measures that address the traceability of the AI system during its entire lifecycle?

The lifecycle of an AI system is intricate, and traceability plays a pivotal role at every stage. From the initial data gathering to the final decision output, each step should be rigorously documented. This includes the processes of data labeling and the specific algorithms employed. Such comprehensive documentation ensures that if an AI system's decision is called into question, there's a clear record of how the decision was derived.

For high-risk AI, this is a legal requirement: technical documentation must be drawn up prior to release of the AI system, and be of such quality that compliance with the Act's requirements can be easily demonstrated. This documentation should include at the very least:

1. The AI system's intended purpose, the person/s developing the system the date and the version of the system;
2. How the AI system interacts or can be used to interact with hardware or software that is not part of the AI system itself, where applicable;
3. The versions of relevant software or firmware and any requirement related to version update;
4. The description of all forms in which the AI system is placed on the market or put into service;
5. The description of hardware on which the AI system is intended to run;
6. Where the AI system is a component of products, photographs or illustrations showing external features, marking and internal layout of those products;
7. Instructions of use for the user and, where applicable, installation instructions;

Input data quality

ALTAI

D1a. Did you put in place measures to continuously assess the quality of the input data to the AI system?

Of particular attention is a continuous assessment of the quality of input data. It's not just about having vast amounts of data; the quality of this data is equally crucial. Regular automated quality assessments can help in identifying issues like missing values, data gaps, breaks in data supply, or even instances where the data is erroneous or mismatched in format. For instance, consider sensor calibration, a process that refines sensor performance by rectifying inaccuracies in sensor outputs. In legal processes, an issue may be that feedback on decisions is not obtained until years later, when an appeal or objection is finally confirmed by the courts.

Several best practices in this area include:

1. **Data Validation Frameworks:** Frameworks like TensorFlow Data Validation and the Python Pandera library generate descriptive statistics, detect anomalies, and ensure data consistency from datasets. Dataset versioning tools like DVC, MLflow, and LakeFS allow organisations to track changes in datasets over time, align model versions with specific data snapshots, and reproduce training conditions during audits or post-market investigations. Refer to chapter 6 on data governance.
2. **Monitoring Data Drift:** Data drift (also known as covariate shift) is the concept that distribution of values within features may shift over time.[5] For instance, a spam

detection filter trained on email from the previous decade will not perform well on today's AI-crafted unsolicited e-mail.

❸ **Data Versioning:** Data versioning permits keeping track of different versions of datasets. This allows for reproducibility and easier identification of when and how data might have changed. Such a process would typically be part of what's called MLOps or machine learning operations.[6]

❹ **Regular Data Audits:** A periodic review of data sources and collection methods to ensure they remain relevant and reliable. We refer back to chapter 5 on data governance.

❺ **Data Annotation Quality Control:** For AI systems that require annotated data (supervised and semi-supervised AI), quality control measures should be in place for the annotation process. This could include periodic reviews, inter-annotator agreement checks, or using multiple annotators for the same data. As with item 3, this would be part of a good MLOps setup.

❻ **Historical Data Backtesting:** Regular testing of the AI system historical data may reveal inconsistencies, indicating the need for further steps to address quality and relevance of the input data.

❼ **Documentation and Metadata:** Comprehensive documentation of data sources, collection methods, preprocessing steps, and any known issues or limitations can help in quickly identifying potential data quality concerns.

Tracing back decisions

ALTAI

D1b. Can you trace back which data was used by the AI system to make a certain decision(s) or recommendation(s)?

D1c. Can you trace back which AI model or rules led to the decision(s) or recommendation(s) of the AI system?

Traceability doesn't end with input data. It's equally vital to be able to trace back the specific data that influenced a particular AI decision or recommendation. This involves understanding which AI model or rules were at play and how they interacted with the data to arrive at a conclusion. This is particularly relevant under the AI Act as well as the GDPR, both of which require explainability of decisions – which just isn't possible if the data and model used is not available in the form used at the time.

The work of Mora-Cantallops et al provides a good review of available tooling for traceability, including backtracing of data.[7] A common measure is to attach metadata referring to source data to every step of the AI process. Unfortunately, this requires a lot of work as most AI pipelines are not designed by default to accommodate this, and may even discard the metadata as irrelevant noise.

A related best practice again is versioning: ensure that the introduction of new datasets or models is clearly separate from earlier models, using names or sequence numbers to allow easy identification. Using the practices from the previous subsection, maintain logs that record every decision made by the AI, the data it used, the model version, and other relevant parameters. This creates a clear audit trail.

Output quality

ALTAI

D1d. Did you put in place measures to continuously assess the quality of the output(s) of the AI system?

Furthermore, the output of the AI system should also be under continuous scrutiny. Standard automated assessments can ensure that prediction scores align with expected ranges and anomalies in outputs are promptly detected. If an anomaly is identified, it's crucial to reevaluate the input data that led to the unexpected output, ensuring that the system remains reliable and trustworthy.

Three key principles underscore the importance of output quality:[8]

1. **Repeatability:** This refers to the ability to obtain consistent measurements under identical conditions. In the context of AI systems, it implies that an investigator can consistently reproduce a particular prediction or other output, using the same procedures and systems, across multiple trials.
2. **Replicability:** This principle emphasizes that a different group, given the same experimental setup, should be able to achieve results from the AI system with the stated precision.
3. **Reproducibility:** Perhaps the most stringent, this principle dictates that even with a different team and a different experimental setup, consistent measurements should be achievable. For AI systems, it signifies that an independent group should be able to obtain the same results from the same source data, even if they develop their tools and artifacts from scratch.

At the very least, an AI Act-compliant system should follow the principle of repeatability. The other principles are best practices to ensure high quality operations.

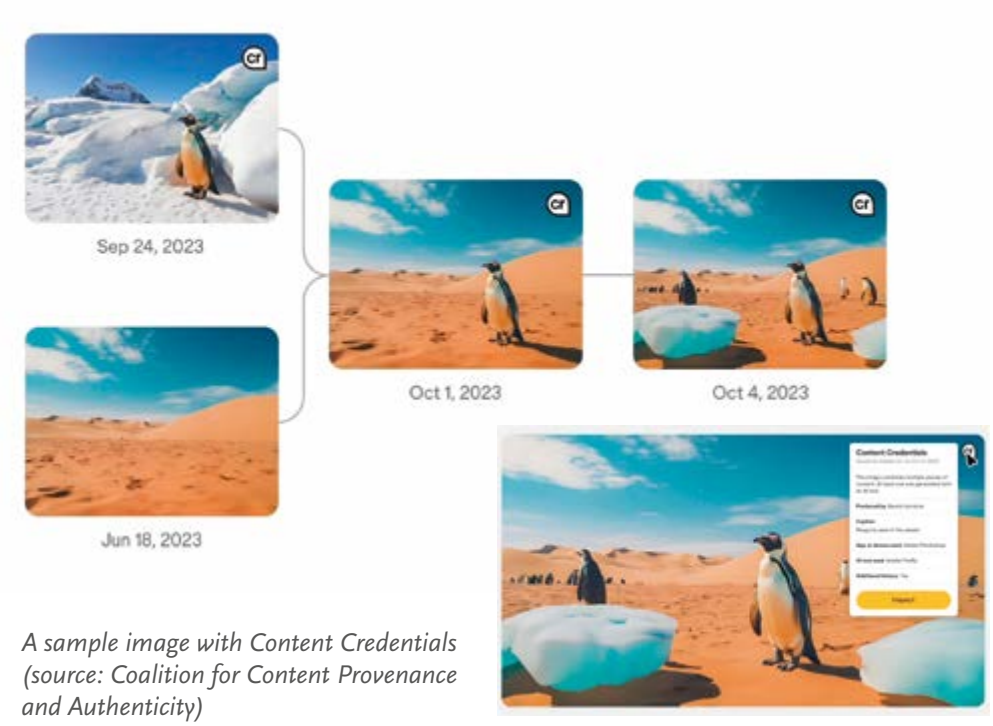
Output of generative AI

A particular concern with AI output is that of generative AI. As the quality of such system rapidly increases, it becomes harder and harder to distinguish AI-generated content and real-world images, audio or movies. This is a cause for concern, especially when dealing with newsworthy events; the issues of fake news, disinformation and so-called deepfakes have been widely discussed.[9] The European lawmakers have chosen to address deepfakes and hard-to-distinguish output of generative AI with a dual

legislative strategy. The AI Act requires providers of such systems to ensure that these outputs are marked in a machine-readable format and detectable as artificially generated or manipulated. Watermarking and other hard-to-remove techniques are recommended solutions.[10] This allows others that intend to use the output to weigh its value and trust. Next, the Digital Services Act (DSA) requires provides of so-called very large online platforms – such as Facebook, Instagram and TikTok – to actively search for these markings and cause appropriate warnings to appear.

The most prominent initiative for AI transparency is the Coalition for Content Provenance and Authenticity (C2PA), an open technical standard backed by more than 1,200 organizations, including Adobe, Microsoft, and Intel. C2PA enables the embedding of tamper-evident metadata directly into digital content, such as images, video, and audio, documenting who created the content, when, and with what tools or modifications. This metadata – referred to as content credentials (CR) – can be cryptographically signed and remains detectable even if attempts are made to edit or remove it. In addition to being embedded in the file, content credentials can also be stored remotely, allowing platforms to verify the authenticity of media even when the local file has been altered.

The below image provides an example. The left side illustrates the operations applied using various image elements, e.g. the desert background insertion and the addition of



A sample image with Content Credentials (source: Coalition for Content Provenance and Authenticity)

blocks of ice. Clicking on the Content Credentials "CR" icon on the right-side image reveals information on the involvement of AI, with the option to retrieve detailed information on the individual elements. This is not to say that the image was manipulated for nefarious purposes; CC is a neutral scheme designed to allow identification of modifications. Human judgment is still required to determine the authenticity of the image and its usefulness in e.g. legal proceedings or information gathering.

Logging practices

ALTAI

D1e. Did you put adequate logging practices in place to record the decision(s) or recommendation(s) of the AI system?

By maintaining a detailed record of the AI system's decisions and recommendations, an AI provider creates a robust framework that not only enhances transparency but also serves as a foundation for future improvements and refinements. The practice of logging is an integral component of software engineering and system administration. When done right, logging provides a clear, chronological account of events, aiding in debugging, monitoring, and understanding the behavior of a system.

Best practices in logging emphasize the importance of clarity, consistency, and relevance.[11] Logs should be clear and concise, avoiding verbosity that can clutter and obfuscate the essential information. Consistency in log format ensures that logs can be easily parsed and analyzed, while relevance ensures that only pertinent information is logged, avoiding the pitfalls of information overload. Furthermore, sensitive information should never be logged, ensuring that user data and system secrets remain secure.

There is no single standard for logging quality, as it depends highly on the application, its intended purpose and expected risks what information should be logged. However, common logging frameworks and libraries are widely available for almost every programming language or development environment. Examples include the log4j Java and Logrus Golang logging frameworks. Other languages, such as the Python language widely used in machine learning, have built-in logging capabilities. Logging frameworks specific to machine learning environments are also available, a common example is TensorBoard.

Logging is not just a good software development practice, it's the law: the AI Act explicitly requires high-risk AI systems to have state of the art logging facilities, enabling the monitoring of operations and events that may pose a risk to fundamental rights. Research on best practices for logging machine learning operations is still ongoing.[12]

Explainability: Making AI understandable

ALTAI

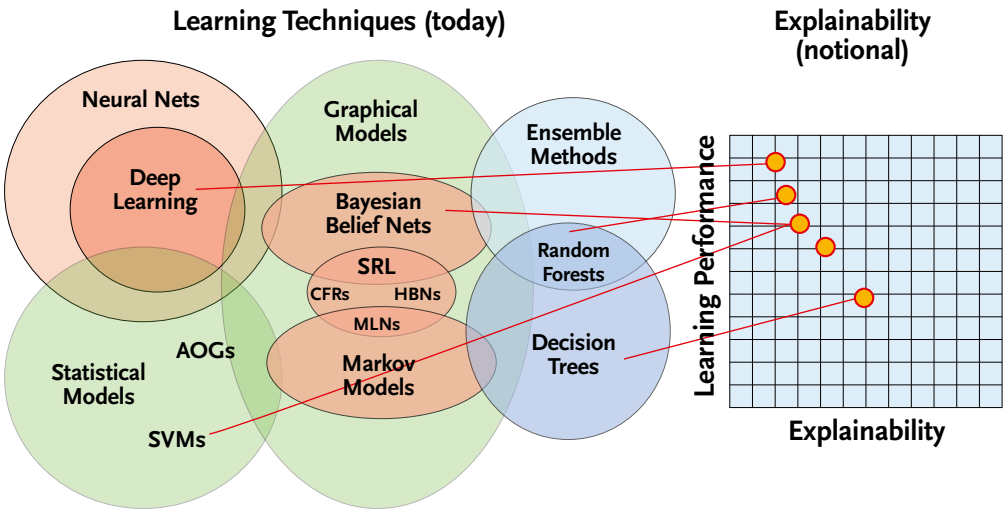
D2. Did you explain the decision(s) of the AI system to the users?

Explainability is often cited as a key factor for adoption of AI systems in a wide range of contexts. This includes the legal context: the GDPR contains a right to obtain "*meaningful information about the logic involved*" – commonly interpreted as a "right to an explanation". Yet, defining what an explanation is remains a still open research question.[13]

Balancing technical explainability and human decisions

Explainability of a machine learning model is usually inverse to its prediction accuracy – the higher the prediction accuracy, the lower the model explainability.[14] This fundamental issue is illustrated in the figure below, which is based on DARPA research. For instance, decision trees have an excellent degree of explainability but exhibit the worst prediction accuracy among the listed learning techniques. In the other extreme, Deep Learning methods are better in predictive capacity than any other learning methods but they are least likely to be explicable.

The requirement for explainability has been criticized as being overly limiting to innovation. For instance, in the field of credit scoring in the USA there is a stringent legal requirement that credit risk predictions must be maximally accurate and come with the highest transparency and auditability. This leaves developers with only very simple predictive models, which actually harms the quality of predictions.[15] In the



Explainability versus quality (source: Xu et al., 2019)

medical field, it is generally accepted that evidence is needed *that* medication or treatment works but not *why* that medication works. From this, arguably explainability in medical AI would be unnecessary if the effectiveness of the AI can be proven through standard clinical trials.[16] A more general criticism is that an AI outcome may be based on millions of features taken in combination. It is then simply unrealistic to expect a comprehensible and easy to follow explanation in layman's terms.

Both the AI Act and the GDPR require AI deployers to provide adequate explanations with AI-made decisions. But, other than generally nothing such decisions should be clear and contain the main elements, they do not clarify what constitutes an 'adequate' explanation. Here we see a difference between fields. In general, the point of an explanation is to teach on the underlying cause of a phenomenon.[17] But in law, the main point of an explanation is to enable an objection or appeal by laying out the main elements and applied argumentation.[18] And here we run into a fundamental problem with AI systems: they do not reason or apply facts to derive from a general rule a specific outcome.

Explaining Deep Learning

As noted in chapter 6, datasets used to train machine learning (ML) models contain a variety of features. For motor vehicles, features could for instance be number of wheels, engine power, maximum speed, and so on. In a loan application systems, factors such as past creditworthiness, savings and checking account balance would be features from which the system derives its decision-making model.

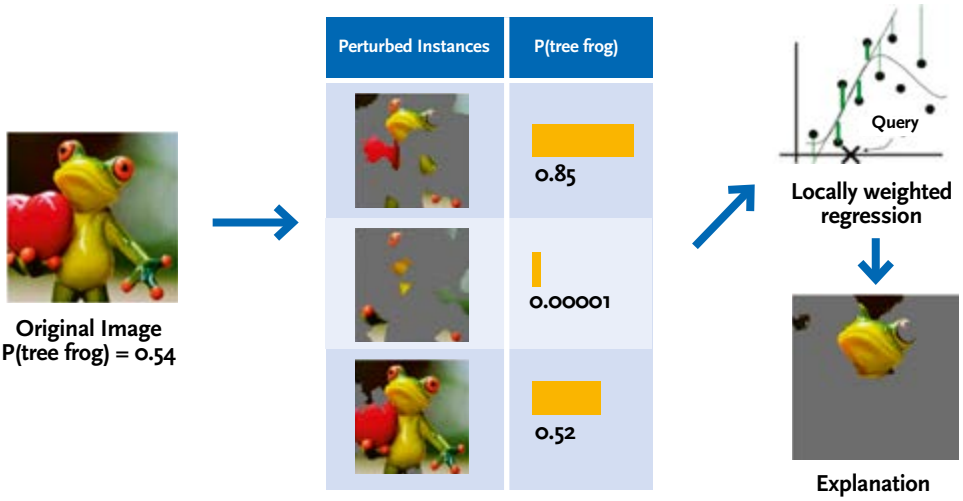
However, it's key to remember that a ML system does not derive a rule-based decision-making model: it creates a probabilistic model based on patterns and correlations found in the data. This means that while the system might recognize that individuals with higher savings balances are generally more creditworthy, this is not used as a *rule* or weighted against other rules. Instead, the data items in the set are divided into groups based on patterns spotted in the data – correlations, in statistical terms.

"Correlation does not imply causation" is an old maxim in statistics. The fact that two features go together does not mean that one is the cause of the other, or that the two even have a mutual relationship. For example, in the well-known *German Credit* dataset used for teaching basic statistics, one feature of loan applicant data relates to whether or not the applicant has a landline telephone.[19] This feature correlates well with the expectation that the loan will be paid back, but no human at a bank would ever cite this fact as a relevant reason to approve or reject a loan. Nevertheless, an AI system trained on this dataset will certainly spot the correlation and thus be more likely to approve loans to people having landlines.

XAI: Breaking the black box

The term "black box" and its synonym *opacity* have become synonymous with models that operate without offering clear insights into their decision-making processes. It is clear that under the AI Act, a high-risk AI system cannot be a black box. In the literature, several approaches have been proposed to add transparency to machine learning systems. This field of study has become known as "XAI", short for "eXplainable AI".[20]

1. **Model Simplification:** Using simpler models that are inherently more interpretable, such as linear regression or decision trees. While these might not always achieve the same accuracy as complex models, they offer more transparency. In case the actual AI system's and the simpler model's outcomes deviate significantly, human intervention may be needed to resolve the matter.
2. **Feature Visualization:** Techniques that visualize the importance of different features in the model's decision-making process. However, as mentioned earlier, feature importance doesn't always equate to human reasoning. This approach therefore is only usable when feature importance is sufficient for transparency purposes.
3. **Post-hoc Explanations:** Much research has been done on methods that provide explanations after the AI has made a decision. Techniques like LIME (Local Interpretable Model-agnostic Explanations) or SHAP (SHapley Additive exPlanations) fall into this category. The general approach is the identification of key features and their relevance to the output categorisation. For instance, in text sentiment classification LIME may highlight words that correlate strongly with the reported sentiment and words that strongly point to another sentiment.



A demonstration of LIME in use on an image classifier (source: Ribeiro et al., 2016)

The concept of XAI was put on the map by LIME. This system was introduced in 2016 as a general technique to add explanations to individual predictions.[21] The below image visually illustrates how LIME adds explanations to an image classifier. The original image on the left was classified as a 'frog'. LIME creates variations of the image called "perturbed instances", as seen in the middle. For each of these instances the original system is asked to make a new classification with probability. The relevance of each section is thus derived and a weighted conclusion is made. In this case, the conclusion is that the frog face in top middle of the image was the most pertinent factor in making the determination.

More recent XAI techniques focus on aligning explanation methods with the internal architecture of neural networks. For example, Integrated Gradients helps identify which parts of an input (e.g., image pixels or text tokens) most influenced a model's decision.[22] In transformer-based models, attention rollout and visual attention maps reveal how the model distributes its focus, offering powerful tools for tracing reasoning pathways.

Post-hoc explanations, by design, offer insights into the AI's decision-making process after the fact. This approach is particularly beneficial in scenarios where the AI model is complex and not inherently interpretable. By translating the AI's decisions into natural language, post-hoc explanations make the outcomes more accessible and relatable to users, regardless of their technical expertise. This fosters a sense of transparency and can alleviate concerns or skepticism users might have about the AI's decisions.

However, there are several risks and downsides to relying solely on post-hoc explanations:

1. **Over-reliance on Simplicity:** While simplifying data-driven decisions into natural language is beneficial, there's a risk of oversimplifying the explanation to the point where it no longer accurately represents the AI's decision process. This can lead to misunderstandings or misconceptions about how the AI operates.
2. **Potential for Misleading Explanations:** Post-hoc explanations are generated based on the model's outputs, but these might not always capture the true underlying reasons for a decision. There's a risk that the explanation provided might be plausible but not entirely accurate.
3. **Template Structure:** Post-hoc explanations may follow a stringent format, and thus be very similar in structure and approach. This may cause the reader to gloss over the explanation and fail to notice subtle differences in a particular case.
4. **Temporal Limitations:** Since post-hoc explanations are generated after the AI's decision, they might not be timely enough for scenarios where real-time understanding is crucial. This can be problematic in situations where immediate human intervention is required based on the AI's decision.

❺ **Overconfidence in AI Decisions:** While post-hoc explanations can enhance trust, they might also lead users to place undue confidence in the AI's decisions without critically evaluating the rationale provided. A well-known phenomenon for instance is that detailed textual explanations are seen as a sign of *competence of the AI system,* and therefore accepted at face value.

A more general point of criticism against post-hoc explanations is that often the AI system is used in an adversarial context. Usually, a user who requests a decision or evaluation from the AI system has different interests than the provider of that system, such as with a loan applicant and a bank, a shopper and mall security or a patient and a healthcare provider. In each of these scenarios, the user seeks a favorable outcome, while the provider aims for accuracy, efficiency, and risk mitigation. What a provider would consider adequate in this situation is quite different from what the user would want. This problem is hard to solve.

An alternative approach may be to offer explanations in the form of counterfactuals: "You were denied a loan because your annual income was €30,000 and your age is below 27. If your income had been €45,000 and your credit score above 0.65, you would have been offered a loan." This type of explanation avoids translating specific and perhaps hard-to-grasp features and their interaction into a complete explanation.[23] The counterfactual indicates what went wrong, what was missing and how this could (in theory) be rectified. Key to a good counterfactual is to present the smallest change that would lead to the desired result. An alternative to counterfactuals is a contrastive explanation, which addresses the question: "Why this outcome rather than that one?".[24] Both show insight in key factors underlying the decision. A common limitation is that the explanations given provide no justification why the factors used should be the decisive ones. They are – simply because the data says so.

User surveys

ALTAI

D3. Do you continuously survey the users if they understand the decision(s) of the AI system?

Continuously surveying users to gauge their understanding of the AI system's decisions presents an alternative approach to post-hoc explanations. The focus on real-time feedback helps to ensure that the AI system's decisions are not just technically correct but also intuitively understood by its users. If users consistently indicate that they do not understand or trust the AI's decisions, it's a clear sign that the system's explanations, whether they are post-hoc or real-time, are not effective.

By continuously collecting feedback, organizations can dynamically adjust and refine the AI's explanatory mechanisms, ensuring that they align more closely with human intuition and reasoning.For instance, in a medical diagnosis AI tool, if patients consistently indicate through surveys that they don't understand why a particular treatment was recommended, the healthcare provider can take steps to improve the system's explanatory capabilities or provide additional training to medical staff to better communicate the AI's decisions.

However, there are challenges to this approach. Continuously surveying users can be seen as intrusive or burdensome, leading to survey fatigue. There's also the risk of receiving skewed feedback if only a subset of users, perhaps those with strong positive or negative feelings, choose to respond. Despite these challenges, when implemented thoughtfully, continuous user surveys can serve as a valuable tool in the quest for more understandable and trustworthy AI systems.

Transparency and automated decision-making

The concept of human agency was presented in chapter 4 as a key concept to understanding when AI can be considered trustworthy. While humans do not generally mind automation of tasks (even decision-making tasks), their sense of self-worth and agency is reduced if they lack meaningful control over the action as a whole. This is core to understanding the various legal and ethical objections against automated decision-making. Automated decision-making has been referred to as an "erosion of human dignity".[25]

Types of decision-making

The process of automated decision-making can be divided into four main categories based on the nature of human involvement:[26]

1. **Supporting:** Providing information to a human decision maker to help them make a decision about a case, but where they are just one source of information amongst others under consideration. This is clearly not 'solely' automated decision-making, unless the human were to blindly adopt the recommendation from the supportive AI system. A simple trick to avoid the latter is to not include a specific conclusion.
2. **Positive Triaging:** New cases are profiled and categorized. The categorization determines the future decision pathway that the case continues along: some category can be automatically processed, while other(s) require human review. The former category would typically constitute the 'positive', 'easy' or 'standard' cases, such as

loan applications that clearly meet all requirements or an airplane passenger raising no red flags. Passengers that do raise flags are then selected for human review.

❸ **Full Triaging**: Similar to positive triaging, but with only the unclear, low confidence or otherwise hard to automate cases being presented to humans. In the loan application example, both the applications that clearly meet the requirements and those that are clearly deficient are automatically processed.

❹ **Summarizing**: Here the human decision-making occurs prior to the automated processing. One or more human decisions or assessments are recorded as structured data, and that data is summarized or consolidated automatically to generate an overall score or assessment which is used to make a decision. For instance, a teacher can review essays and assign scores on topics such as spelling, consistency and quality of work. The automated system would use the teacher's scores to create a full and motivated review for the student, including calculating the final grade.

Addressing automated decision-making in law

In legal circles, the concept of automated decision-making is often connected to the GDPR, which contains a long-standing provision against automated decision-making. According to its article 22, humans "shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her." Note the word 'solely': if there is a certain level of human oversight, the decision is not 'solely' automated. However, this must be more than a token human gesture such as rubberstamping the computer's recommendation. The European Data Protection Board uses the term "meaningful human involvement".[27]

Further, the decision must be produce legal effects or otherwise significant effects. A legal effect is one that affects one's legal rights, e.g. by cancelling a contract, denying access to elections or refusing admission to enter the country. This definition is somewhat muddy – is refusing to enter into a contract a legal effect, considering that no one has a right to demand a contract? For this type of case, the second prong of the article is relevant: denying a contract application has a similarly significant effect to cancelling a contract and is therefore also covered by article 22. The effect must be significant and long-lasting. For instance, denying an employment opportunity is covered, but a profile-based decision to hide certain content on social networks is not.

When a person is subjected to an automated decision, article 15 of the GDPR entitles them to an explanation on the procedure and principles actually applied in the case, not merely a general statement of how the algorithm is supposed to work. The explanation must be sufficient to enable the affected person to file an appeal or objection, or at the very least have the decision reviewed by a human person.

Unclarity over the scope of the GDPR's right to an explanation has led the drafters of the AI Act to add a similar right (article 86) in the case high-risk AI systems take decisions affecting humans. A key difference is that under the AI Act, any decision "on the basis of the output from a high-risk AI system" qualifies, while the GDPR only demands explanations for "solely automated" decisions. In contrast, the scope is very much limited: only decisions within high-risk use cases of Annex III (see chapter 2) require explanations, while under the GDPR any decision based on personal data is sufficient.

The ethical limits of explanations

While offering a rationale for an automated decision may satisfy transparency obligations, it does not resolve the deeper problem: whether the underlying logic is justifiable in ethical or legal terms.[28] An explanation does not cleanse a flawed or discriminatory system. Indeed, a system may offer a plausible reason for its decision ("You were denied because your postal code correlates with higher risk"), while relying on criteria that are socially prejudicial or legally impermissible.

This reveals a fundamental asymmetry: algorithmic systems often deliver outputs without reasons — just patterns, correlations, or inferred signals. Presenting these as if they were human-like justifications can obscure their normative emptiness. Worse, it can create a false sense of legitimacy, especially when paired with technical jargon or selective saliency maps. The result is a form of epistemic injustice: affected individuals are asked to trust decisions they cannot meaningfully contest, shouldering the burden of understanding systems they never chose, in terms they never consented to.[29]

If left unchallenged, this risks turning explanation into a smokescreen, a veneer of due process over systems that reproduce bias, amplify inequality, or quietly encode private interests. Transparency is vital, but it must be tied to accountability (chapter 10). The right to know why is only meaningful if paired with the right to demand that decisions be fair, grounded, and open to challenge. Providing insights in system operations through transparency and explanations thus are necessary steps – but only the first ones. Rather than treating explicability as a retrofit, systems should be architected to support meaningful challenge, legal reasoning, and normative scrutiny from the outset – be contestable by design.[30]

Communication: Bridging the gap between AI and users

The final aspect of transparency to be discussed is communication. AI systems are very novel and users are generally expected to be unfamiliar with them. The onus therefore is on AI producers, distributors and deployers to provide the information users need to be able to work with the systems.

Recognizing the AI Interface

ALTAI

D4. In cases of interactive AI systems, do you communicate to users that they are interacting with an AI system instead of a human?

Distinguishing between human and machine interactions is crucial, as already discussed in earlier chapters. This distinction ensures that users can set appropriate expectations regarding the nature and limitations of the responses they receive. For instance, while chatting with customer support, knowing that one is communicating with an AI can help users tailor their queries more precisely, understanding that they might not pick up on nuances or emotions in the same way a human would. This type of transparency is an explicit requirement (article 50) of the AI Act. Chatbots and the like must always be clear on their artificial status.

Clarity on purpose and criteria

ALTAI

D5. Did you establish mechanisms to inform users about the purpose, criteria, and limitations of the decision(s) generated by the AI system?

Every AI system is designed with a specific purpose in mind, and it operates based on certain criteria. Users have a right to understand the primary objective of the AI they're interacting with. For example, a recommendation engine on a streaming platform aims to suggest shows or movies based on a user's viewing history and preferences. Being aware of this can help users appreciate why certain content is being recommended to them and can guide them in refining their preferences if needed.

The AI Act puts the 'intended purpose' of AI systems front and center. Any testing and risk evaluation is done with the intended purpose in mind, and the intended purpose must be disclosed in the documentation (see next section). The Act defines this term as "the use for which an AI system is intended by the provider, including the specific context and conditions of use" and confirms that what it says on the box is what the system is supposed to do – the small print cannot change that.

Highlighting the benefits

ALTAI

D5a. Did you communicate the benefits of the AI system to users?

While AI systems bring numerous advantages, it's essential to communicate these benefits clearly to the users. This not only fosters trust but also ensures that users can make the most of the technology. For instance, a predictive text feature can speed up typing and improve accuracy, but users might only fully utilize it if they understand its potential benefits.

One of the most effective ways to communicate the benefits of an AI system is through interactive demonstrations. By allowing users to engage with the AI in real-time, they can directly experience its advantages. For instance, a photo editing software with AI-enhanced features could offer a side-by-side comparison where users can upload an image and see the before and after effects of the AI's enhancements. This hands-on approach can be more impactful than merely reading about the benefits.

Addressing technical limitations

ALTAI

D5b. Did you communicate the technical limitations and potential risks of the AI system to users, such as its level of accuracy and/ or error rates?

As part of the transparency obligations, the AI Act requires producers of high-risk AI to disclose the levels of accuracy and the relevant accuracy metrics of their systems in documentation. The previous chapter goes into more detail on the various options and trade-offs. What's important is to also look beyond the metrics as such: a 98% accuracy in voice recognition software is great, but does that apply across all languages or only American-accent English?

Training and disclaimers

ALTAI

D5c. Did you provide appropriate training material and disclaimers to users on how to adequately use the AI system?

Providing users with adequate training materials and clear disclaimers can ensure that they know how to interact with the AI optimally. For instance, a medical diagnostic AI tool might be highly efficient, but without proper training on inputting patient data correctly, its predictions might be off. The system can also caution users against over-relying on the AI and remind them to use their judgment in critical situations.

The term 'disclaimer' may call up the image of small print written by a lawyer to address a variety of situations. It is well-known that such disclaimers have little to no practical effect. Disclaimers should be user-centric: they should help the user if and when they approach problematic situations. Sensitive content warnings on social networks are a prime example of user-centric disclaimers. These alert users about potentially distressing content, such as graphic images or discussions on sensitive topics. By giving users a heads-up, they can choose whether or not to engage with the content, ensuring they have control over their digital experience. In an AI system, such a warning can serve to alert the user that he or she is about to generate graphic or otherwise sensitive outputs.

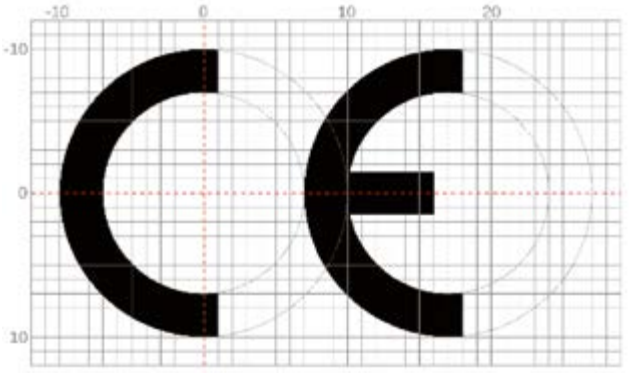
Another effective use of disclaimers is the mandatory acknowledgment before accessing specific features. For example, software applications with advanced functionalities might present a warning that requires users to click "I Understand" before proceeding. This ensures that users are aware of the potential risks or complexities of the feature they're about to use. In an AI system this could be used when the user requests an action or output that would exceed the intended use of the AI system.

Moreover, tutorial pop-ups in applications can also be seen as a form of disclaimer. They guide users through new features, ensuring they understand the functionality and potential implications before diving in. If the user goes too fast, the action could be refused and the user redirected to an earlier step or simpler procedure. This proactive approach not only educates users but also reduces the likelihood of misuse or unintended actions.

The AI Act even goes one step further: documentation and training must also consider "reasonably foreseeable misuse" of the system, and users must be aware of the consequences of such misuse. Overreliance (see chapter 4) is a clear example: relying on AI decisions or recommendations instead of applying human judgment. Another example is an image generating AI tool that is misused by a paying customer to create deepfakes with the intent of passing them off as real. The system should have safeguards against this, and the documentation should address the undesirability of this type of use. (This is not the same as the terms of use stating this is prohibited.)

Clarity out of the box: the CE logo

The AI Act is part of the New Legislative Framework of the EU, which aims to improve the internal market and boost the quality of conformity assessments. The old Conformité Européenne (CE) marking, which translates to "European Conformity" is given a new life. Since the 1980s, this distinctive symbol signifies a product's compliance with European Union (EU) health, safety, and environmental protection standards and thus is an early example of transparency, the subject of this chapter.



The AI Act applies the CE logo into the domain of AI systems. Providers of high-risk AI systems must undertake a conformity assessment, compile a technical dossier with all relevant documentation, and sign an EU Declaration of Conformity. Only after this step can the CE logo be affixed to the system – and without the logo, the system may not be deployed in the European Economic Area. This way, the European citizens' trust in the CE marking can also apply to the use of high-risk AI systems. The advantage of the marking for producers, importers and distributors of AI systems is that they enjoy the freedom of movement; Member States should not create unjustified obstacles to the placing on the market or putting into service of high-risk AI systems that comply with the AI Act and bear the CE marking.

Key takeaways

We have seen the paramount importance of transparency in fostering trust and ethical interactions. We've dissected the intricacies of traceability, the nuances of explainability, and the profound implications of automated decision-making. Furthermore, the emphasis on clear communication highlighted the necessity of distinguishing between human and AI interactions, ensuring users are well-informed and can set appropriate expectations. These takeaways are foundational in understanding the ethical deployment of AI systems. Next, we'll look at how AI can be fair, diverse, and non-discriminatory.

Fostering Fairness, Diversity, and Non-Discrimination

The pursuit of fairness, diversity, and non-discrimination has become increasingly central to discussions about artificial intelligence and algorithmic decision-making. While technology promises efficiency and objectivity, reality has shown that AI can also amplify existing societal prejudices, leading to unequal and unfair outcomes. This chapter explores how fairness, diversity, and non-discrimination principles can be embedded into AI development and deployment processes. It addresses both the theoretical foundations and practical steps needed to mitigate bias, ensure equitable representation, and build systems that respect diverse populations.

Introduction to fairness, diversity, and non-discrimination in AI

Fairness is paramount in the development and deployment of AI systems. While interpretations of fairness vary, it encompasses both substantive and procedural dimensions. Substantively, fairness mandates an equitable distribution of benefits and costs, ensuring freedom from bias, discrimination, and stigmatization. Procedurally, decisions by AI should be contestable, with clear accountability and explainable processes. Let's start with a closer look at these concepts.

The imperative of fairness

In the context of AI, the term 'fairness' has its roots in statistics. Statistical fairness exists when an AI system's outcomes do not disproportionately favor or disadvantage any subgroup of a dataset based on attributes that are independent of the selection criterion.[1] For example, in the context of creditworthiness a statistical model would be called unfair if its predictions were influenced by a person's gender or ethnicity rather than solely their financial history and current financial status, as such features have no relevance in such an analysis. However, in a study analyzing the prevalence

By the end of this chapter, you'll be able to ...

- Understand and implement fairness measures.
- Prioritize accessibility and inclusivity.
- Apply the principles of Universal Design in creating inclusive and accessible AI systems.
- Engage and collaborate with stakeholders.

of certain genetic diseases, ethnicity would be a relevant selection criterion due to the known genetic variations and predispositions among different populations.

In society, fairness is a fundamental right: the equitable treatment of all individuals regardless of their inherent or acquired characteristics, ensuring that every person has equal access to opportunities and resources, and is protected from unjust discrimination, bias, or prejudice in all spheres of life. In Europe, these values are reflected in art. 21 of the Charter and art. 14 of the European Convention on Human Rights. The AI Act and the Guidelines recognize fairness as a building block for trustworthy AI: a fair AI system fosters trust among its users, promotes equitable outcomes, and paves the way for a more inclusive digital future.

The concept of 'bias'

Discussions on fairness go hand in hand with the concept of 'bias'. Again, this term originates from the field of statistics, where it refers to systematic errors in estimates or inferences.[2] Bias occurs when a model's predictions consistently deviate from the actual values it's trying to estimate. For instance, if a machine learning model consistently underestimates the wear on a car tire, it exhibits bias. Common forms of bias in statistical models include:

1. **Measurement bias:** Arises when data is consistently measured incorrectly. This could be due to faulty equipment or human error. In the car tire example, this could be a defective sensor that fails to count one in ten wheel revolutions.
2. **Sampling bias:** Occurs when the sample collected is not representative of the entire population. For instance, if the car tire data is obtained from measurements in taxis, this would not be representative of the entire population of car drivers. To avoid this, data should be collected from a diverse range of drivers and vehicle types.
3. **Model bias:** Introduced when the assumptions made by a model do not align with the real-world data. For example, if the car tire model assumes that tire wear is solely based on distance driven and neglects factors like road conditions or driving style, it would exhibit model bias by not accurately predicting wear in scenarios with frequent hard braking or rough terrains.
4. **Historical bias:** This occurs when the data used to train an AI model reflects past prejudices, inequalities, or unfair practices that existed in society. For example, historically, tire salesmen may have sold only smaller, less durable tires to women, believing them to be less knowledgeable about cars or driving shorter distances. An AI system trained on this historical sales data might continue to recommend such tires to female customers, despite these preferences being based on obsolete gender stereotypes rather than actual driving habits or needs.

5. **Confirmation bias:** This occurs when data is selectively chosen, consciously or unconsciously, to confirm a pre-existing belief or hypothesis. For instance, if a researcher believes that car owners would obviously swap tires according to the manufacturer's recommended driving distance, they might be inclined to ignore measurements that exceed such recommendation.
6. **Selection bias:** This type of bias arises when the data used to train a model is not representative of the broader population. For example, if a tire wear estimation model is trained only on data from cars driven in summer conditions, it might not perform well when estimating wear for tires used in winter conditions. Selection bias often occurs due to the way data is collected or chosen for analysis, rather than just being about representativeness.

The difference between sampling and selection bias in particular can be subtle. As an example, if the car tire model only sampled cars from a single neighborhood, it would exhibit a sampling bias if that neighborhood has unique driving conditions not representative of the entire city. In contrast, the model would exhibit selection bias if all car owners in the city were invited to participate in the data gathering, but only infrequent drivers chose to respond and participate (e.g. because the date and time of the data gathering was particularly inconvenient for frequent drivers). In this case, the voluntary participation of a specific group with distinct driving habits would skew the results, making them not representative of the average car owner in the city.

Bias and discrimination in AI and algorithms

In everyday language, 'bias' is often equated with 'discrimination', which refers to the unjust or prejudicial treatment of individuals based on attributes like race, age, or gender. It's important to note that the term 'discrimination' carries a heavy connotation, suggesting malicious intent or hate towards certain groups. Usage of this term can therefore derail discussions, as the vast majority of AI developers and businesspeople might feel unjustly accused and act defensively.[3] The term 'bias' is perceived as more neutral.

That's not to say AI systems never exhibit forms of bias that negatively affect groups of people.[4] In fact the opposite is true. Gender biases have been identified in word embeddings in natural language processing systems, e.g. associating male terms more closely with career-oriented words and female terms with family-oriented words.[5] Facial recognition systems often have difficulty recognizing non-white skin, notably Google Photos mistakenly labeling African American faces as "gorillas".[6] Algorithmically chosen job advertisements often present higher-paying jobs to men rather than women.[7] And the list goes on.

Bias in AI systems often emerges unintentionally, embedded through training data or algorithmic choices. In a 2024 study, a generative AI system tasked with drafting job interview evaluations consistently produced more favourable descriptions and ratings for candidates whose names and language patterns were associated with ethnic majority groups.[8] Conversely, individuals with names commonly linked to minority ethnic or cultural backgrounds received evaluations containing subtly negative phrasing and fewer affirmations of competence.

Next to the general prohibitions on discrimination, fairness in data processing in particular is enshrined in article 8 of the European Charter of Fundamental Rights and article 5 of the GDPR.[9] The AI Act specifically requires providers of high-risk AI systems to examine their data sets for possible biases that may "have a negative impact on fundamental rights or lead to discrimination" and to apply appropriate mitigation measures (as discussed in chapter 7).

Inclusive engineering

Inclusivity is about ensuring that all individuals, regardless of their background, identity, or ability, are included, considered, and actively sought to participate in a process or system. In the context of AI or product design, inclusivity means designing systems, interfaces, or products that are usable by as many people as possible, including those with disabilities or those from diverse backgrounds.

While AI systems are rooted in technical intricacies, it's essential to recognize that addressing fairness and inclusivity goes beyond mere code adjustments. Simply put, ensuring diversity and inclusion isn't just about tweaking algorithms or removing certain data columns like gender or ethnic background. It's a holistic endeavor that encompasses understanding societal nuances, ethical implications, and the broader human experience.

Relying solely on an engineering-centric approach can inadvertently simplify and even misrepresent the profound ethical challenges at hand.[10] Engineers, designers, and product managers, while experts in their domains, might sometimes view challenges through a lens that emphasizes abstraction and formalization. This can lead to a scenario where vital concepts like diversity and inclusion are treated as mere technical parameters or checkboxes to be met. Moreover, an overly technical mindset might shift the ethical onus away from AI developers and providers. Instead of taking full ethical responsibility, they might perceive their role as merely offering a set of tools or configurations for users to navigate. However, the AI Act underscores that every entity in the AI value chain, be it producer, distributor, or deployer, bears the responsibility for the ethical and legal integrity of the AI systems. A strategic approach is therefore necessary.

Establishing strategies and procedures to avoid bias

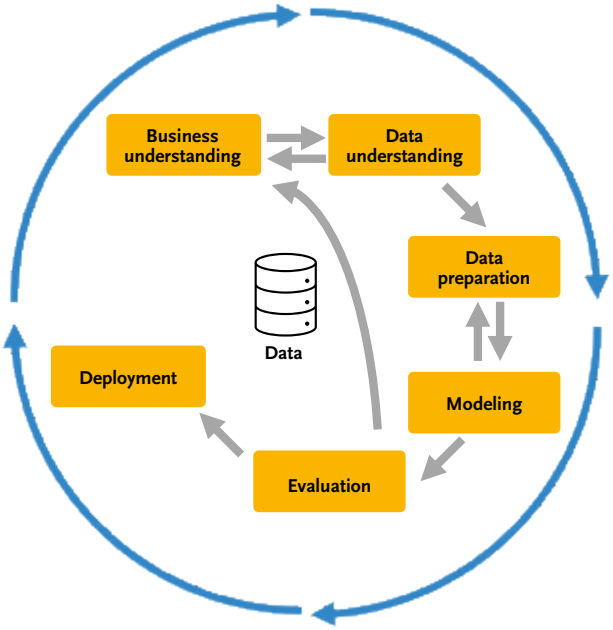
ALTAI

E1. Did you establish a strategy or a set of procedures to avoid creating or reinforcing unfair bias in the AI system, both regarding the use of input data as well as for the algorithm design?

To ensure that AI systems are both fair and effective, it's crucial to adopt a comprehensive strategy that addresses potential biases at every stage of the model development process. The CRISP-DM (Cross-Industry Standard Process for Data Mining) process model provides a structured framework that can be adapted to this end.[11] While the model dates from the late 1990s, its approach is still very much sound, assuming the project is goal-directed and process-driven. CRISP-DM remains one of the clearest and most widely used process models in AI development.

However, alternative frameworks such as Data Science Trajectories (DST) offer a more iterative and risk-aware structure, especially useful for public sector or high-accountability domains.[12] DST adds reflections on data source selection, stakeholder goals, value assumptions, and communication of results, encouraging practitioners to revisit decisions iteratively: not just how a model is built, but why, for whom, and with what embedded assumptions. These exploratory layers helps ensure that fairness risks are surfaced early, dataset representativeness is contextually evaluated, and that transparency and stakeholder relevance are preserved throughout the lifecycle.

The CRISP-DM model can be visualized as follows:

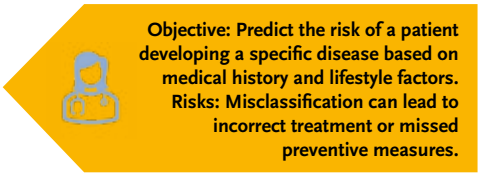


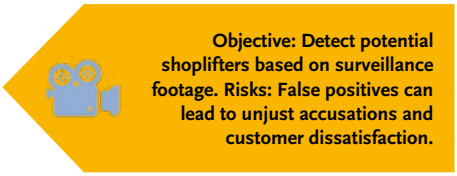
Step 1: Business understanding

The process starts with a clear definition of the business objectives for the AI model. This involves understanding the potential risks associated with biases, especially in sectors where decisions can have significant real-world consequences, such as healthcare or finance. By identifying potential areas of concern early on, teams can be more vigilant in subsequent stages.

Fairness issues must be considered already at the start of a project, because the way in which the business problem is framed often determines whether an unfair or discriminatory model is likely to emerge. For example, using "propensity to buy" in retail to determine what neighborhoods should receive price discounts may create discriminatory outcomes even if the algorithm itself is technically neutral.

We will use two running examples in this section to illustrate application of the process. The first focuses on predicting patient risk in healthcare, and the second on improving anti-theft surveillance in a supermarket. The business objectives and risks could be formulated as follows:





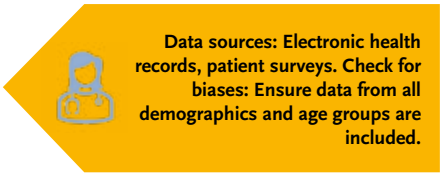
Step 2: Data understanding

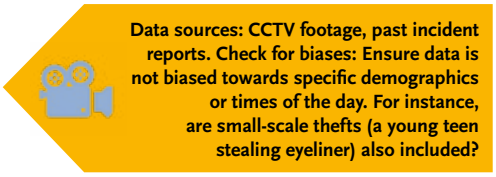
The second step focuses on the nature of the input data. Knowing how data has been obtained, what biases may be in the sample and how particular features correlate with protected characteristics is key. This step must also address the risk of inherited bias, where upstream systems (e.g. GPAI) have introduced hidden patterns that propagate downstream. Bias may be inherent in the input data, such as over-representation of men in management, or in the method of gathering or classifying data.

Some work has been done on creating statistical techniques to identify bias. The key principle is to generate a large number of outcomes, which are then correlated with gender or other features where bias is suspected. If a correlation can be found, bias should be investigated further. This does require processing of gender or other special categories of personal data, which triggers stringent GDPR requirements. However, the AI Act provides an explicit legal basis for this particular processing, as it is in the public interest to have bias-free AI systems. A DPIA with clear justification on why the special personal data is strictly needed is a key requirement.

Specific statistical techniques commonly used for this purpose are:

1. **Disparate Impact Analysis:** This technique measures the difference in outcomes between different groups. For instance, in a hiring algorithm, if one demographic group has a significantly lower selection rate than another, it may indicate potential bias. The "four-fifths rule" is a common threshold used in disparate impact analysis: if the selection rate for a particular group is less than 80% (or four-fifths) of the selection rate for the group with the highest selection rate, it may be evidence of potential bias.
2. **Odds Ratio:** The odds ratio is a measure used to compare the odds of an event occurring in one group to the odds of it occurring in another group. In the context of bias detection, if the odds ratio is significantly different from 1, it might indicate a potential bias between the two groups. For example, in a credit scoring model, if the odds of being classified as a high-risk borrower are three times higher for one demographic group compared to another, it suggests potential bias.
3. **Residual Analysis:** In regression models, residuals (the differences between observed and predicted values) can be analyzed to detect potential biases. If residuals are systematically higher or lower for specific groups, it may indicate that the model is not fitting well for those groups, suggesting potential bias. For instance, if a model consistently underestimates the performance of certain demographic groups, it might be biased against them.



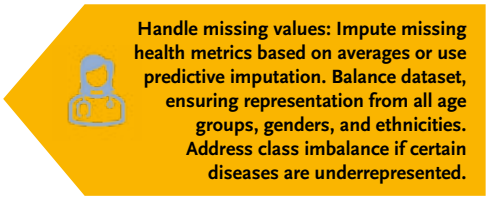


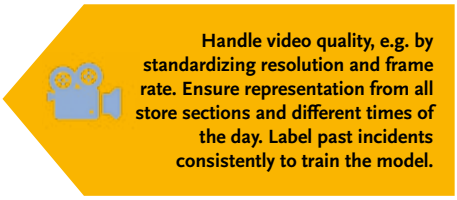
Step 3: Data preparation

Once the data is understood, it must be prepared for modeling. This step has grown significantly in importance due to the central role that data plays in shaping fairness outcomes. Data preparation needs particular focus on bias mitigation (in the statistical sense), requiring techniques such as reweighing, optimized preprocessing, and embedding debiasing. This step is also where inherited bias, such as from general-purpose AI systems (GPAI), must be evaluated and mitigated. If synthetic or augmented datasets are used, they must be assessed for demographic balance and traceability to avoid amplifying unseen bias patterns. Practices such as Datasheets for Datasets or Dataset Nutrition Labels are increasingly standard, helping teams record provenance, collection context, and known limitations.[13]

Typical fairness-aware data preparation includes:

1. Cleaning and normalization: Removing inconsistent or erroneous entries, harmonizing formats and units.
2. Feature engineering with bias sensitivity: Identifying proxies for sensitive attributes and testing them for indirect discrimination effects.
3. Handling imbalanced data: Applying resampling, augmentation, or weighting to improve group representation during modeling.
4. Bias-aware synthetic data use: Using generative tools to augment underrepresented groups or test edge-case behaviors without compromising real user data.



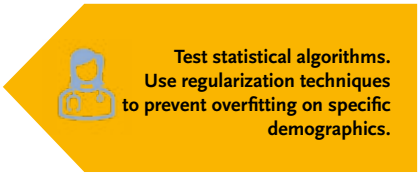


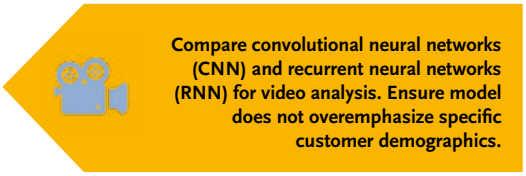
Step 4: Modeling

The choice of a machine learning algorithm plays a pivotal role in determining the fairness of the resulting model. Different algorithms have varying sensitivities to biases present in the data, and their inherent design can also introduce or amplify biases. Evaluating multiple algorithms thus is a clear necessity.

Today, few AI developers create their own models or algorithms from scratch. Many foundation models or GPAI models are on the market to quickly start with classification, content generation, recommendation and so on. As training a model from scratch requires significant computational resources and time, using foundation models can drastically reduce these costs.

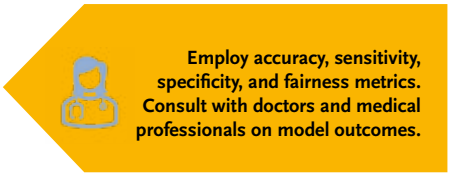
However, foundation models are typically trained on general datasets that aim to capture a broad spectrum of information. To tailor these models to specific applications, developers often fine-tune them using their own datasets. This allows for customization and can improve performance for the task at hand. Yet, the general data they are trained on might carry inherent biases from the sources they were collected from. There are also the issues of transparency and explainability to consider. Foundation models, especially deep learning models, can be complex and not easily interpretable. This can pose challenges in applications where understanding the decision-making process of the model is crucial. At the same time, creating an own model from scratch may exhibit similar problems.

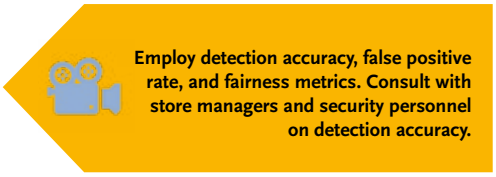




Step 5: Evaluation

After modeling, the AI system should be rigorously evaluated for both performance and fairness. This involves using metrics that specifically measure bias and fairness, in addition to traditional performance metrics. Stakeholder consultations are key to gather feedback on potential real-world implications. Both are used to iteratively refine the model based on evaluation results.

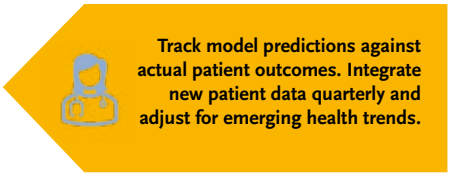




Step 6: Deployment

Once deployed, continuous monitoring of the AI system is crucial. Real-world data can change, and new biases can emerge. Strategies for this phase include setting up automated bias detection mechanisms, regularly updating the model with fresh data to ensure it remains relevant and fair and engaging in periodic stakeholder consultations to stay informed about changing contexts and needs.

Bias can also arise in this step, for instance if the AI system receives systematically different input data in production compared to the training data, or if the model performance deteriorates over time for specific groups. This is why the AI Act prescribes Post-Market Monitoring or PMM (see chapter 3). Organizations must implement practical pipelines to detect fairness drift, track input-output changes over time, and log group-specific model behavior for auditability.



Ensuring diversity and representativeness

ALTAI

E2. Did you consider diversity and representativeness of end-users and/or subjects in the data?

E2a. Did you test for specific target groups or problematic use cases?

E2b. Did you research and use publicly available technical tools, that are state-of-the-art, to improve your understanding of the data, model, and performance?

E2c. Did you assess and put in place processes to test and monitor for potential biases during the entire lifecycle of the AI system (e.g., biases due to possible limitations stemming from the composition of the used data sets (lack of diversity, non-representativeness)?

E2d. Where relevant, did you consider diversity and representativeness of end-users and or subjects in the data?

Considering the diversity and representativeness of end-users and subjects in the data is crucial. An AI system trained on a diverse and representative dataset is more likely to produce fair and unbiased results. This involves ensuring that the data encompasses various demographics, backgrounds, and scenarios to reflect the real-world diversity. Unfortunately, research on methodologies for data collection and annotation with the purpose of ensuring diversity and representativeness is scarce.

While early efforts drew inspiration from archival data ethics (e.g. transparency, mission statements, consent), current state-of-the-art methods additionally involve proactive demographic auditing, use of crowdsourcing with representation goals, and formal protocols for data source mapping. Today's datasets are expected to capture sociocultural, geographic, and behavioral diversity that reflects real-world populations, especially in high-impact domains such as employment, education, or health.

Projects can improve representativeness through:[14]

1. **Demographic distribution analysis:** Comparing dataset makeup against real-world population statistics using tools like Fairkit or Aequitas.
2. **Targeted data supplementation:** Actively sourcing data from underrepresented groups or edge-case scenarios.
3. **Bias simulation:** Testing how exclusion of certain subgroups impacts model outcomes across use cases.
4. **GPAI-aware review:** When using general-purpose AI systems, deployers should trace the representativeness limitations of upstream training data and apply correction or caveat strategies accordingly.

Modern AI systems, especially those deployed in dynamic environments or built on GPAI, must be monitored for performance drift, subgroup disparities, and emergent harms. This requires integrating bias evaluation checkpoints into post-market monitoring (PMM) and ongoing quality assurance cycles. Monitoring must go beyond

static metrics to include temporal and contextual fairness, recognizing that bias can emerge over time or under specific real-world conditions.

An oft-cited concern is that testing for bias necessarily involves the use of so-called special personal data, data on sensitive aspects such as ethnic origin, sexual orientation, religious beliefs or political leanings. The GDPR generally bans the collection and use of such data, which is for good reasons but does hamper bias analysis. The AI Act's drafters included a special provision under which providers (not deployers) may exceptionally use such sensitive data for the specific purpose of bias mitigation. Its relation to the GDPR is somewhat unclear, and the many requirements in the provision make it somewhat ill-suited for its purpose.[15]

Education and Awareness Initiatives

ALTAI

E3. Did you put in place educational and awareness initiatives to help AI designers and AI developers be more aware of the possible bias they can inject in designing and developing the AI system? The design and development stages of AI system?

The design and development phases of AI systems are particularly prone to the unintentional introduction of bias. These stages often reflect the assumptions, experiences, and cultural positioning of the designers and developers, factors that can silently shape how fairness and discrimination risks manifest in AI behaviour.

Under the AI Act (see Chapter 3), organizations developing or deploying AI are expected to cultivate AI literacy among staff, particularly those involved in the various stages of system lifecycle. AI literacy thus here extends to an understanding of fairness risks, legal and ethical implications, and the societal context in which systems operate.

Effective programs go beyond generic training. They include structured workshops on identifying cognitive and systemic bias in design decisions, using fairness tools and metrics (e.g. AIF360, Fairlearn) and navigating the trade-offs between performance and non-discrimination goals. Awareness initiatives can also include regular seminars, interdisciplinary case reviews, and internal audit simulations where teams reflect on past projects. These practices foster a culture of reflection and accountability, crucial for building trustworthy AI.

Mechanisms for flagging issues

ALTAI

E4. Did you ensure a mechanism that allows for the flagging of issues related to bias, discrimination or poor performance of the AI system?

E4a. Did you establish clear steps and ways of communicating on how and to whom such issues can be raised?

E4b. Did you identify the subjects that could potentially be (in)directly affected by the AI system, in addition to the (end-)users and/or subjects?

A robust mechanism should be established that allows stakeholders, users, and even the broader public to flag issues related to bias, discrimination, or subpar performance of the AI system. For instance:

❶ **Bias Detection Toolkits:** Implementing software toolkits that allow users to test the AI system's outputs against various demographic groups. By comparing results, these toolkits can highlight disparities in outcomes, indicating potential biases. An example could be a toolkit that analyzes facial recognition software's accuracy across different ethnicities, pointing out if certain groups are misidentified more frequently than others.

❷ **Feedback Portals with Bias Categories:** Creating dedicated online feedback portals where users can report perceived biases. These portals can have specific categories or tags related to common bias concerns, such as gender bias, racial bias, or age-related bias. By categorizing feedback, it becomes easier to analyze and address specific areas of concern.

❸ **Stakeholder Review Panels:** Organizing periodic review panels consisting of diverse stakeholders who evaluate the AI system's decisions in real-world scenarios. These panels can assess if the AI system's decisions are favoring certain groups over others and provide recommendations for adjustments. For example, a panel reviewing a hiring AI tool can assess if the tool is consistently ranking certain demographic groups lower than others for no justifiable reason.

❹ **Clear Guidance:** Providing step-by-step guidance on the reporting process ensures that individuals know exactly how and to whom they can voice their concerns.

Beyond the immediate users of the AI system, it's essential to recognize that the ripple effects of AI decisions can impact a broader audience. The AI Act recognizes this by defining "affected person" as anyone who could be indirectly affected by the AI system's outcomes. This includes not just end-users but also individuals or groups who might be influenced by the decisions, predictions, or actions of the AI system. For instance, in autonomous car design any person participating in traffic is affected by the car's decisions, not just the driver or passengers.

Any affected person should be able to flag issues if he or she is affected by the system. This is not trivial: often, feedback mechanisms are only available to customers. Outside parties often face impregnable walls before being able to find contact points, or are directed to general customer service representatives with no particular connection to the AI developers.

A solution could be found by borrowing inspiration from the field of IT security. In this realm, companies often employ 'bug bounty' programs, where they incentivize external experts to find and report vulnerabilities in their systems. Just as these programs reward individuals for identifying weak spots that could be exploited by malicious actors, imagine a 'bias bounty' system for AI. In this system, organizations would reward individuals or groups who identify and report biases in AI models and algorithms. Just as bug bounty programs have become a cornerstone in ensuring software security, a bias bounty system could be a revolutionary step in ensuring fairness and transparency in AI. By tapping into the collective expertise of a diverse community, organizations can benefit from a wide range of perspectives, ensuring that AI systems are as unbiased and equitable as possible.

Defining and measuring fairness

ALTAI

E5. Is your definition of fairness commonly used and implemented in any phase of the process of setting up the AI system?

E5a. Did you consider other definitions of fairness before choosing this one?

E5b. Did you consult with the impacted communities about the correct definition of fairness, i.e. representatives of elderly persons or persons with disabilities?

E5c. Did you ensure a quantitative analysis or metrics to measure and test the applied definition of fairness?

E5d. Did you establish mechanisms to ensure fairness in your AI system?

Ensuring fairness is not just about selecting a particular metric or definition but involves a holistic approach that considers various stakeholders, especially those who might be impacted by the system's decisions. Settling on a particular definition helps align those involved in the development.

There are many definitions of fairness in AI, each tailored to different contexts, goals, and trade-offs. No single metric can fully capture what is "fair" across all situations. Instead, practitioners must choose fairness metrics that align with the use case, the societal or legal risk, and the system's intended impact. The table below summarizes commonly used fairness metrics, outlining what each measures, when it is typically applied, and its key limitations.

Metric	What It Measures	Use When...	Limitations or Trade-Offs	Best Use Case
Equal Opportunity	Whether true positive rates are equal across groups (error-based)	Screening for eligibility (e.g. hiring, loan approval)	Can reduce overall accuracy or disproportionately affect other groups	When false positives and negatives have different impacts
Equalized Odds	Whether both true positive and false positive rates are equal (error-based)	Balancing fairness with predictive consistency	Hard to optimize; often conflicts with other goals	When identifying positive cases correctly is crucial (e.g., disease detection)
Predictive Parity	Whether precision (positive predictive value) is equal across groups (error-based)	When outcomes are high-stakes and errors carry cost (e.g. recidivism prediction)	Conflicts with equal opportunity when base rates differ	When identifying negative cases correctly is crucial (e.g., security screening)
Calibration by Group	Whether predicted probabilities reflect true outcome likelihood per group (calibration-based)	GPAI outputs or risk scoring (e.g. insurance, health)	Does not ensure equal opportunity or selection rates	When group representation is the primary goal
Disparate Impact Ratio	The ratio of positive outcome rates between protected and unprotected groups (regulatory parity)	Legal or compliance contexts (e.g. hiring audits, Title VII)	Requires threshold justification; doesn't indicate cause	When prediction calibration is critical (e.g., risk assessment)

Applying multiple fairness metrics concurrently can provide a more comprehensive view of an AI system's fairness, but doing so involves trade-offs. Metrics often reflect different conceptions of fairness, and their requirements can directly conflict.

Consider a fraud detection system, where the positive class (fraud) is rare and the cost of false negatives (rejected payments) is high. Equal Opportunity would ensure that true positive rates (i.e. successful fraud catches) are equal across groups, such as individual versus corporate

accounts. In contrast, Demographic Parity might distort model behavior by requiring equal fraud flagging rates across groups, even if fraud prevalence differs.

In domains like hiring or loan screening (where decisions are consequential and populations have roughly balanced qualifications) Equal Opportunity or Equalized Odds can be used. Equal Opportunity focuses on true positives (e.g. qualified candidates being selected), while Equalized Odds additionally considers false positives. These metrics help ensure that systemic biases don't mask themselves in uneven error rates, especially in sensitive decision-making contexts.

In content recommendation systems (e.g., news feeds, music platforms), the fairness challenge often lies in representation rather than accuracy. Popular or mainstream content tends to dominate, while niche content can be algorithmically marginalized. Demographic Parity can help ensure equal exposure across content categories or demographic profiles, but if outcomes like click-through rates differ naturally, Calibration by Group may offer a more realistic fairness check – confirming whether predicted probabilities align across groups. Equalized Odds might be too rigid in such cases, leading to reinforcement of existing imbalances without addressing root representational skew.

Accessibility and Universal Design

ALTAI

E6. Did you ensure that the AI system corresponds to the variety of preferences and abilities in society?

It is no longer sufficient for AI to be merely functional; it must be inclusive, catering to the diverse tapestry of human experiences and needs. This does not just mean in the dataset; users from a variety of backgrounds should gain the same experience and use from the AI system. So let's examine some of the best practices for ensuring diversity, universal access and usability.

Ensuring accessibility in AI system design

If AI systems are not designed with a broad spectrum of users in mind, they risk excluding or disadvantaging certain segments of the population. For instance, individuals with visual impairments might struggle with AI-driven visual content platforms that lack alternative text descriptions or voice-over functionalities. Similarly, those with hearing impairments could be left out if voice-activated AI assistants don't provide visual feedback or captioning. Physical disabilities might limit the use of touch or gesture-based interfaces, and neurodivergent individuals might process information differently, requiring a more straightforward and less metaphorical interaction.

Designing a system usable by all has been a challenge throughout the history of computer systems, and designing AI systems has been no different. The field of Human-computer interaction (HCI) has been focusing on the design of computer technology and, in particular, the interaction between humans (the users) and computers for a long time. Already in the 2000s, the field focused on product design incorporating dilemmas on automation versus human control, expanding in the early 2010s towards incorporating the impact machine learning.

In the European Union, accessibility is not merely a commendable design principle; it's a binding legal requirement. The European Accessibility Act (Directive 2019/882) mandates a wide range of products and services to be accessible by all. This includes ticketing machines, check-in machines, computers, telephones, televisions, banking services, e-books, and e-commerce.[16] The EAA emphasizes the importance of harmonized standards for products and services. A key standard is EN 301 549, enabling organizations to measure the accessibility of websites, electronic documents and non-web software such as native mobile apps, against documented success criteria for all people, including those with disabilities.[17] For public sector bodies, stricter rules follow from the earlier Web Accessibility Directive (2016/2101). Compliance with these directives is a legal requirements for all providers of high-risk AI systems.

Making user interfaces usable by all

ALTAI

E7. Did you assess whether the AI system's user interface is usable by those with special needs or disabilities or those at risk of exclusion?

E7a. Did you ensure that information about, and the AI system's user interface of, the AI system is accessible and usable also to users of assistive technologies (such as screen readers)?

E7b. Did you involve or consult with end-users or subjects in need for assistive technology during the planning and development phase of the AI system?

As with all computer systems, the user interface (UI) of an AI system serves as the primary bridge between the technology and its users. Ensuring that this bridge is sturdy, welcoming, and inclusive is paramount. Every individual, regardless of their abilities or needs, should be able to navigate, understand, and benefit from AI systems. Accessibility by all has always been a key issue, whose importance everyone agrees with; unfortunately industry – at large – has not embraced proactive approaches.[18] A potential reason is that although the total number of persons having specific accessibility needs is large, each individual need represents only a small portion of the population, making it impractical to design everything so that it is accessible by everyone regardless of their limitations.[19] This however is a misconception: there is no need for a "one-size-fits-all" design.

Assistive technologies have long acted as vital tools in bridging the gap between AI systems and users with specific needs. Screen readers, for example, vocalize digital content for those with visual impairments. Voice recognition software allows users to command and control systems without the need for traditional input methods. Braille displays transform digital text into tactile braille characters. They however often served as afterthoughts: a UI is designed for on-screen interaction with a mouse, and then voice interaction is added to accommodate impaired users.

Today, multimodal interaction presents a different paradigm. Systems should be designed to be interacted with in multiple ways. For example, consider a smart home assistant device. A user might use voice to ask the device to play their favorite song or provide the weather update, tap on the device's screen to adjust settings or browse through options, wave their hand in front of the device to skip a song or mute an alarm and receive information through on-screen visuals, like a dynamic weather map or a video clip.

To ensure AI systems are accessible to all:

- Developers should adhere to established accessibility standards, such as EN 301 549 which is in practice mandatory to establish EAA compliance.[20]
- Systems should be designed with adaptive interfaces, offering multiple modes of interaction.
- Regular testing with assistive technologies should be conducted to identify and rectify potential compatibility issues.
- End-users, especially those relying on assistive technologies, should be actively involved to acquire insights that can shape the trajectory of AI development.

Universal Design principles in AI development

ALTAI

E8. Did you ensure that Universal Design principles are taken into account during every step of the planning and development process, if applicable?

Universal Design (UD) is an approach that seeks to create products and environments that are inherently accessible to both people with disabilities and those without. The concept is also known under names such as inclusive design, Design for All and especially in an EU context as eInclusion and eAccessibility.[21] The seven principles of Universal Design are:

1. **Equitable Use:** The design should be useful and marketable to people with diverse abilities. For example, a voice-activated virtual assistant that can understand and respond to a wide range of accents, speech patterns, and languages, ensuring that users from different linguistic backgrounds can use it effectively.

❷ **Flexibility in Use:** The design should accommodate a wide range of individual preferences and abilities. For example, an AI-driven learning platform that offers content in multiple formats – videos, text, audio descriptions, and interactive simulations – allowing users to choose the mode that suits their learning style best.

❸ **Simple and Intuitive Use:** Use of the design should be easy to understand, regardless of the user's experience, knowledge, language skills, or concentration level. An AI system should thus uses simple, clear language and visual aids to explain complex terms or procedures appropriate to the target audience.

❹ **Perceptible Information:** The design should communicate necessary information effectively to the user, regardless of ambient conditions or the user's sensory abilities. For instance, an AI traffic navigation system should provide both auditory directions and visual maps, ensuring that drivers can receive information through their preferred sensory channel.

❺ **Tolerance for Error:** The design should minimize hazards and the adverse consequences of accidental or unintended actions. An AI photo editing tool for example should permit multiple undo levels and confirms potentially irreversible actions, reducing the chance of users making unintended changes.

❻ **Low Physical Effort:** The design should be usable comfortably and with a minimum of fatigue. The use of voice commands is a common way to achieve this, e.g. in an AI-powered home automation system.

❼ **Size and Space for Approach and Use:** Appropriate size and space should be provided for approach, reach, manipulation, and use, regardless of the user's body size, posture, or mobility. For example, an AI-driven virtual reality game that adjusts its interface and controls based on the user's height, arm length, and mobility, ensuring an immersive experience for all.

Assessing AI system impact on end-users

ALTAI

E9. Did you take the impact of the AI system on the potential end-users and/or subjects into account?

E9a. Did you assess whether the team involved in building the AI system engaged with the possible target end-users and/or subjects?

E9b. Did you assess whether there could be groups who might be disproportionately affected by the outcomes of the AI system?

E9c. Did you assess the risk of the possible unfairness of the system onto the end-user's or subject's communities?

The development and deployment of AI systems have profound implications for end-users and subjects. As already stressed elsewhere in this book, this requires a holistic approach to AI development, involving not just the technical team but also the potential users of the system. It's essential to ask:

- Were the target end-users consulted during the development phase?
- Did the development team actively seek feedback from a diverse set of potential users?
- How were the insights and concerns of these users addressed in the final product?

For instance, an AI system designed for elderly care should involve feedback from seniors, caregivers, and medical professionals to ensure it meets the unique needs and challenges of its target audience. The strategies above to mitigate bias should also go a long way to address disproportionate impact on specific target groups.

Stakeholder participation

ALTAI

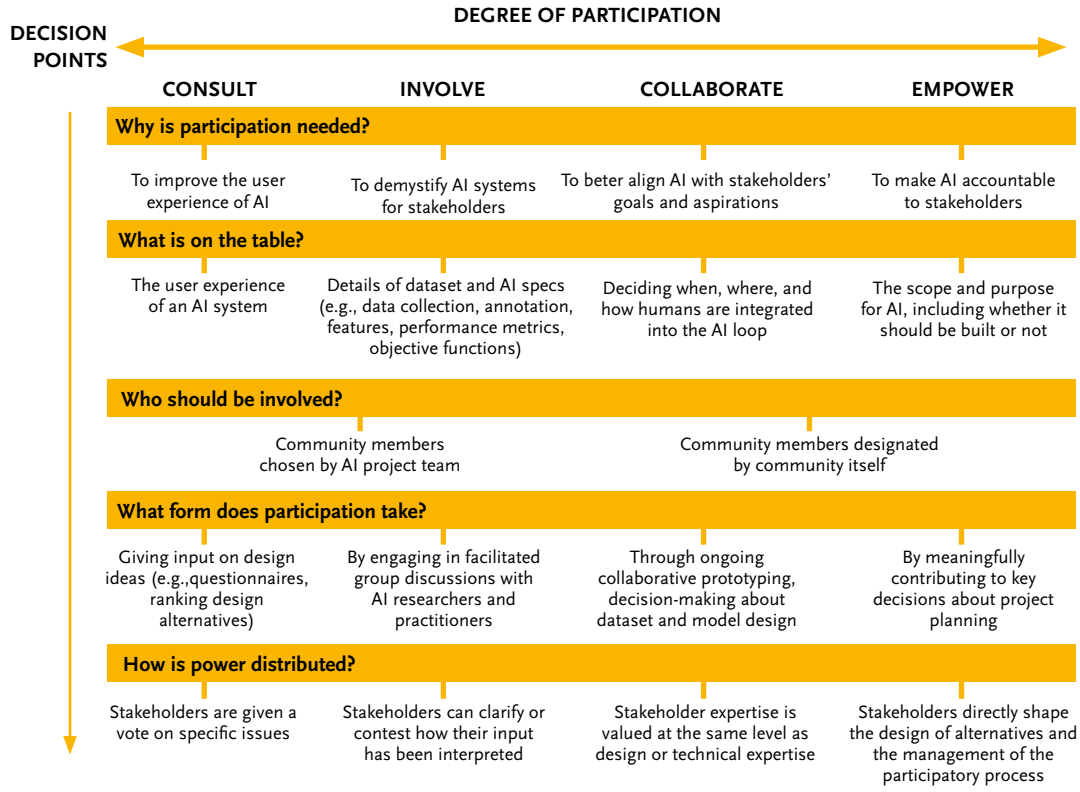
E10. Did you consider a mechanism to include the participation of the widest range of possible stakeholders in the AI system's design and development?

Engaging a diverse range of stakeholders ensures that the AI system is designed with a holistic understanding of its potential impacts and applications. This participation can lead to more robust, ethical, and effective AI solutions. The involvement of key stakeholders is an AI Act requirement for high-risk AI systems, and a recommended option for other AI systems.

Working with stakeholders

In the context of computer systems, a "stakeholder" refers to any individual, group, or organization that has an interest in or is affected by the design, development, deployment, and operation of the system. This can include end-users, developers, business owners, investors, regulatory bodies, and even communities where the system might be deployed. Stakeholders can influence or be influenced by the system's objectives and outcomes.

The field of Human-Computer Interaction (HCI) in particular has always emphasized the importance of understanding and designing for the end-user. An early start was made with the advent of personal computers in the 1980s, which caused a shift towards making systems more user-friendly. Of particular note is the concept of co-design or participatory design, which emphasizes the active involvement of users in the design process (rather than, say, merely giving feedback on the almost-finished prototype). Today, with the rise of complex socio-technical systems, stakeholder participation has become even more critical. Issues like privacy, ethics, and societal impact have brought a wider range of stakeholders into the design and decision-making process.



Tactics for increasing stakeholder participation in designing AI, from consulting to empowering (source: Delgado et al., 2021).

While in the field of AI design there is a strong and growing consensus that end-users and stakeholders should participate, there is still an enormous divergence on what this actually means. Tools such as workshops and feedback rounds are often employed, but without standardization or frameworks often appear to be mere afterthoughts or symbolic gestures.[22] In the literature, five key questions have been identified as key to stimulating participation at the key decision points in an AI design process. They are illustrated in the figure below.[23]

The five questions are:

❶ **Is participation needed?** This requires a good understanding of the necessity and benefits of involving stakeholders in the design process.

❷ **What is on the table?** This step involves determining the specific issues, topics, or decisions that are open for discussion and input.

❸ **Which stakeholders should be involved? This requires identification of all** relevant individuals, groups, or entities that should be part of the process.

4. **What form does their participation take?** The method or approach of involvement is of crucial importance. For instance, using a Social Choice Theory approach, AI practitioners might poll stakeholders and then aggregate their preferences. However, this might not address power imbalances between AI researchers and stakeholders. Alternatively, participatory democracy research suggests more collaborative methods, like bringing stakeholders together to discuss and negotiate design decisions.
5. **How is power distributed among the participating stakeholders and between stakeholders and technology designers/engineers?** Assessing the balance of influence and decision-making authority among all involved parties is key to understanding what the stakeholders actually bring to the table.

There exists a notable disparity between idealistic aspirations and practical limitations. Two primary considerations emerge from this context:

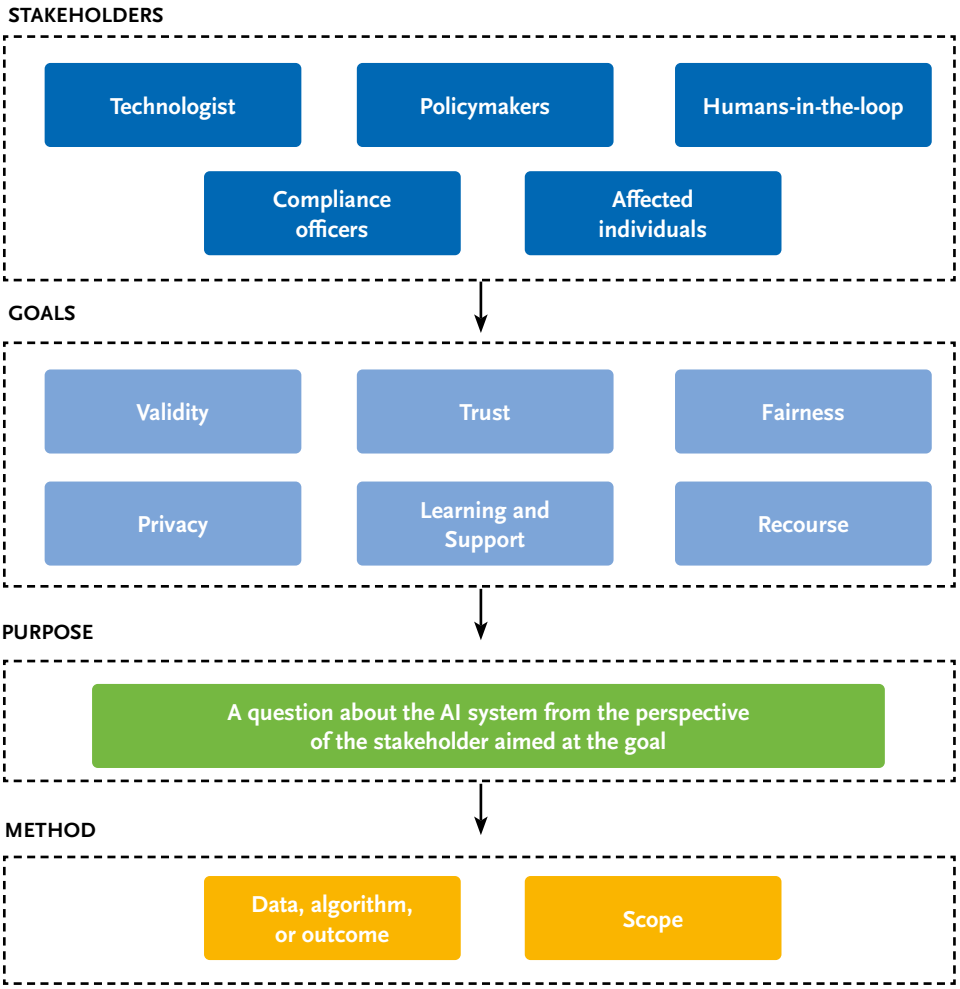
1. What defines the baseline for substantial participation, especially when determining the efficacy of human and algorithmic proxies in representing stakeholder values and preferences?
2. At which juncture does the enhancement of stakeholder involvement yield diminishing benefits?

It's imperative for businesses to address these considerations to optimize stakeholder engagement in AI design and development. A best practice is to conduct a thorough evaluation to establish clear benchmarks for meaningful participation and to determine the optimal level of stakeholder involvement to ensure both efficiency and inclusivity.

Toolkits for participation

To foster creation of ethical or trustworthy AI, a plethora of tools, resources, guides, and kits is available from many sources.[24] Such toolkits often promise to also address participation, but in practice do little more than recommending that stakeholders participate or their voice is heard. For instance, the Microsoft Azure Application Architecture Guide recommends that implementers "seek more information from stakeholders that you identified as potentially experiencing harm". This is little more than hand-washing akin to legal disclaimers.

A useful approach is the stakeholder-first framework by Bell, Nov and Stoyanovich.[25] Their framework (illustrated on the next page) puts the 'technologists' or those who work on the AI system on equal footing with policymakers, users (humans-in-the-loop), compliance officers and affected individuals (the stakeholders, in the terminology of the previous subsection). Assuming algorithmic transparency is intended to improve the understanding of a human stakeholder, AI designers must first consider the stakeholders of the system, before thinking about the system's goals or the technical methods for creating transparency.



A stakeholder-first approach for creating Transparent algorithmic decision-making (source: Bell, Nov and Stoyanovich 2023).

Next, the goals of desired transparency or fairness are defined. These can be put in one of six categories: validity, trust, learning and support, recourse, fairness, and privacy. From here the purpose question – what does the stakeholder want from the AI system aimed at their specific goal – can be formulated. Only then is it possible to select the methods: data, algorithm, outcome and scope.

Key takeaways

Fairness in AI requires attention to both outcomes and processes. Systems must distribute benefits and risks equitably, and their decisions must be accountable and explainable. Using structured development models like CRISP-DM helps address bias across the lifecycle. Diverse, representative datasets and fairness-aware design choices are central to this effort, especially when general-purpose AI systems are involved.

Beyond technical design, inclusivity and accessibility must guide how systems interact with real users. Engaging a diverse range of stakeholders can surface hidden risks and improve both ethical robustness and practical usability. As we turn to the next chapter, we shift focus from fairness within systems to the broader societal and environmental impacts of AI.

9

Societal and Environmental Implications of AI Systems

Artificial Intelligence (AI) has moved beyond its early promise of optimization and efficiency to become a systemic force shaping both the environment and society at large. No longer a matter of speculative ethics, the impact of AI systems is now a regulatory concern, particularly where systems introduce measurable environmental burdens, disrupt labour structures, or influence democratic processes. The AI Act, together with sustainability mandates like the CSRD and CSDDD, positions these societal and environmental effects as legally accountable domains. This chapter explores how organizations must identify, evaluate, and mitigate these risks –not as an afterthought, but as integral components of responsible AI deployment.

Aligning environmental impact with global goals

AILITAI

F1. Are there potential negative impacts of the AI system on the environment?

F1a. Which potential impact(s) do you identify?

From energy use to water stress and e-waste, the environmental footprint of AI has become impossible to ignore. Its impacts stand in direct tension with global commitments under the Sustainable Development Goals (SDGs), as well as Europe's own Green Deal and corporate sustainability requirements. The issue for any organisation therefore is how to identify, quantify, and govern those effects in line with legal and ethical responsibilities.

Environmental impact of AI

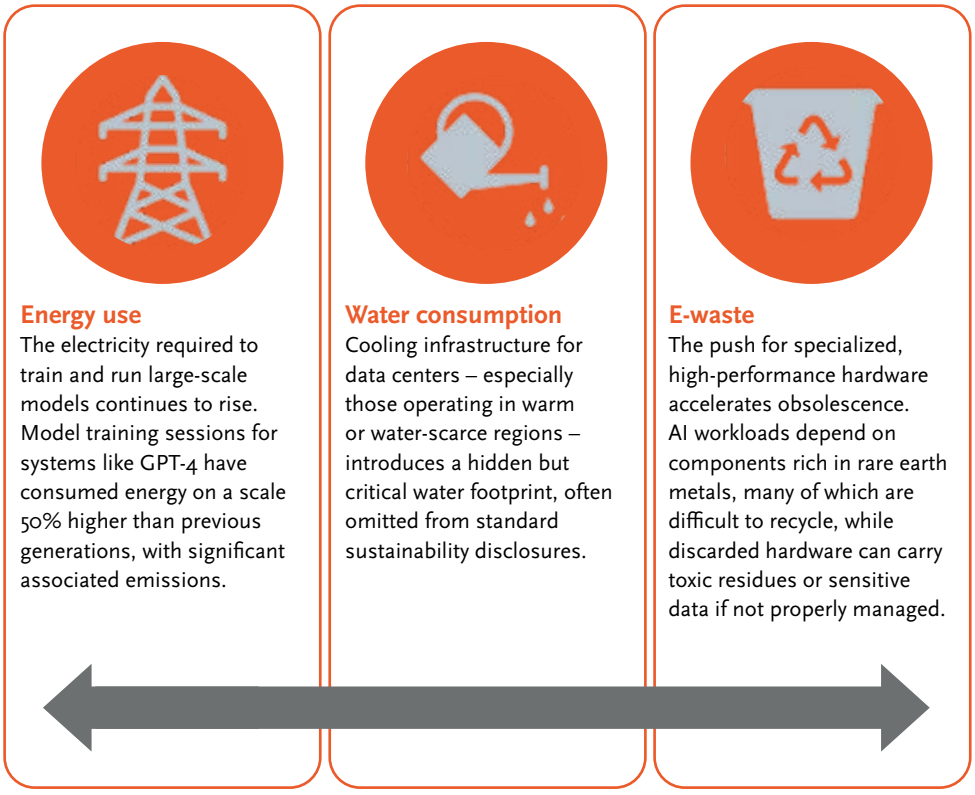
Artificial intelligence systems exert environmental pressure at every stage of their lifecycle, starting from data processing and model training to deployment, storage, and eventual decommissioning. GPAI models in particular have escalated resource demand, consuming vast computational power and accelerating hardware turnover. Recent

By the end of this chapter, you'll be able to ...

- Understand the intersection of AI with global sustainability, societal constructs, and environmental implications.
- Analyze the potential impact of AI on democratic processes and its role within ESG and CSR considerations.
- Develop strategies for mitigating societal and environmental risks.

studies have quantified these impacts in terms accessible to non-technical stakeholders: a single complex prompt can use as much energy as producing three plastic bags; a day's worth of AI assistant usage by one employee can equal the emissions of an 8-kilometer car journey.[1] Scaled across organizations, these effects become significant contributors to carbon emissions and resource depletion.

Three main dimensions define the environmental impact of AI systems:



While reporting of energy use is widespread, the water footprint of AI models is considerable. Training GPT-3 in Microsoft's advanced U.S. data centers consumed an estimated 700,000 liters of freshwater, roughly equivalent to the water needed to manufacture 370 BMW vehicles.[2] The AI sector also drives a significant increase in electronic waste.[3] The pace of innovation in AI hardware leads to rapid obsolescence. Components that are only a few years old can become incompatible with new software architectures or fall short of computational demands. These retired devices often contain hazardous substances like lead, cadmium, and brominated flame retardants. Without proper disposal and recovery processes, such waste can leach into soil and water systems, posing long-term ecological and public health risks.

It's clear that AI systems are no longer neutral bystanders in the sustainability landscape. Their energy demands, water dependencies, and hardware implications place them squarely within the scope of environmental risk assessments and sustainability reporting. This is where the UN's Sustainable Development Goals come in.

The Sustainable Development Goals (SDGs)

The Sustainable Development Goals (SDGs), adopted by the United Nations in 2015, define a comprehensive framework for addressing humanity's most pressing social, economic, and environmental challenges. Spanning objectives such as climate action, clean water, decent work, and reduced inequalities, the SDGs have become a universal language for policy coherence and corporate accountability. They increasingly serve as benchmarks for governmental action and private-sector practice.

Artificial intelligence sits at a complex intersection with these global goals. On the one hand, the resource intensity and environmental footprint of modern AI systems risk undermining progress toward SDG 6 (Clean Water and Sanitation), SDG 12 (Responsible Consumption and Production), and SDG 13 (Climate Action). On the other, AI also presents a unique capacity to accelerate progress in domains like healthcare (SDG 3), sustainable agriculture (SDG 2), education (SDG 4), and climate modeling (SDG 13).

Indeed, many AI applications are already being deployed in support of SDG-aligned initiatives:[4] satellite-based crop yield predictions to improve food security, AI-powered early-warning systems for natural disasters, and intelligent energy systems that optimize renewable integration into the grid. These contributions illustrate that AI is not inherently misaligned with sustainability – it is a tool whose impact depends on design, governance, and deployment context.

The challenge for AI governance professionals is therefore twofold: mitigate environmental harms while maximizing positive contributions. Aligning AI strategies with the SDGs requires a lifecycle approach.

From Social Impact Assessment to governance integration

ALTAI

F2. Where possible, did you establish mechanisms to evaluate the environmental impact of the AI system's development, deployment and/or use (for example, the amount of energy used and carbon emissions)?

F2a. Did you define measures to reduce the environmental impact of the AI system throughout its lifecycle?

Social Impact Assessments (SIAs) were once promoted as a comprehensive tool to evaluate the broader effects of AI systems on society. Their intent – to capture long-term, indirect, or diffuse societal risks – remains relevant. However, in practice, SIAs have struggled with vagueness, lack of enforcement mechanisms, and limited integration into organizational workflows.

By 2025, the focus has shifted toward more concrete and legally grounded instruments. The AI Act's Fundamental Rights Impact Assessment (FRIA, see chapter 3) offers a clearer procedural model with defined scope and obligations. Similarly, the EU's Corporate Sustainability Reporting Directive (CSRD) and Corporate Sustainability Due Diligence Directive (CSDDD) embed societal metrics in annual ESG disclosures, tying them to board-level accountability.

As part of conducting a **FRIA**, deployers must assess whether an AI system's operation causes or contributes to environmental harm. This includes not only direct emissions or water consumption, but also the system's reliance on infrastructure (such as data centers) with high ecological footprints. A compliant FRIA process typically includes:

- Estimating carbon emissions during model training and deployment;
- Measuring energy use per prediction or per transaction;
- Assessing datacenter sourcing (e.g., renewable vs. fossil-powered);
- Evaluating water usage in cooling infrastructure;
- Reviewing hardware turnover and disposal policies.

Complementing this, the **CSRD** and **CSDDD** frameworks create a dual-layered responsibility for organizations. CSRD requires transparent disclosure of environmental impacts across the full value chain, while CSDDD imposes a duty of care to identify, prevent, and mitigate environmental harm. Both include issues arising from AI systems and their supporting infrastructure.

Together, these directives require organizations to:

- Include AI-related energy use and emissions in carbon accounting;
- Report on water usage and cooling practices where AI infrastructure contributes to local water stress;
- Disclose volumes of electronic waste and efforts toward circular economy principles;

- Integrate sustainability criteria into procurement contracts with AI vendors and cloud providers;
- Define and report on measurable targets for reducing AI-related environmental impacts;
- Conduct environmental due diligence on suppliers of AI infrastructure (e.g., chipmakers, data centers);
- Document mitigation efforts and follow-up actions in annual ESG and due diligence reports.

By embedding these measures into FRIA and CSRD workflows, organizations create a repeatable, governance-ready pathway for monitoring and minimizing AI's environmental burden throughout its lifecycle. Practical steps to reduce and govern the environmental impact of AI systems include:

❶ **Select certified low-impact infrastructure** Choose data centers and cloud providers that adhere to the EN 50600 or ISO/IEC 30134 series, disclosing metrics like Power Usage Effectiveness (PUE), Water Usage Effectiveness (WUE), and Carbon Usage Effectiveness (CUE).

❷ **Procure energy-efficient AI hardware** Require AI-relevant hardware (e.g. GPUs, AI accelerators, edge devices) to carry EU Energy Label or Energy Star certification, where applicable. Hardware refresh policies should explicitly favor modular, upgradeable systems with published lifecycle carbon data.

❸ **Track and report model-level emissions** Use standardized tools or internal telemetry to estimate energy consumption and CO_2 equivalents per training run and per inference. These figures should feed directly into Scope 2 and Scope 3 GHG Protocol reporting under CSRD, and inform risk classification in the FRIA.

❹ **Incentivize efficient design through internal pricing** Implement an internal carbon or energy cost for AI workloads, measured in kWh, GPU hours, or CO_2 per inference. This promotes design discipline and aligns AI development incentives with sustainability goals. Consider applying PACT (Partnership for Carbon Transparency) principles when sourcing AI-as-a-service components.

❺ **Offset only as a last resort—with full transparency** Where emissions or resource use cannot be avoided, purchase high-integrity environmental offsets (e.g. Gold Standard, Verra) and disclose them in annual CSRD reports. Offset strategies must never substitute for upstream reduction or responsible infrastructure choices.

AI in the work environment

ALTAI

F3. Does the AI system impact human work and work arrangements?

AI is increasingly embedded in the organization of work, notably in the form of systems that monitor, schedule, evaluate, and direct workers. These systems, often implemented as part of "algorithmic management" platforms, raise new compliance concerns around transparency, consent, fundamental rights. The AI Act here intersects with labour-specific regulations, such as the EU's Platform Work Directive.

Algorithmic management and worker autonomy

ALTAI

F4. Did you pave the way for the introduction of the AI system in your organisation by informing and consulting with impacted workers and their representatives (trade unions, (European) work councils) in advance?

F5. Did you adopt measures to ensure that the impacts of the AI system on human work are well understood?

F5a. Did you ensure that workers understand how the AI system operates, which capabilities it has and which it does not have?

Algorithmic management refers to the use of algorithms in general, and AI in particular, to direct, evaluate, and coordinate workers, all without direct human supervision. These systems are now common across logistics, customer service, manufacturing, and knowledge work, and are deeply embedded in platforms that automate scheduling, productivity tracking, customer rating integration, and even disciplinary decisions.

For workers, the experience of algorithmic management often brings a sense of opacity, surveillance, and detachment from decision-making. Studies across sectors (such as platform work, retail, and call centers) consistently report increased feelings of stress, alienation, and mistrust when automated systems determine schedules, tasks, or performance evaluations.[5] A 2024 Eurofound survey on digital workplace conditions found that employees under algorithmic supervision reported significantly lower job satisfaction and a higher sense of job insecurity compared to those managed by human supervisors. The absence of clear communication, combined with unchallengeable or unexplained decisions, can contribute to a culture of quiet compliance, reduced morale, and erosion of workplace solidarity.

When implemented transparently and with meaningful human oversight, algorithmic management can reduce administrative friction, streamline task allocation, and bring consistency to operational decisions. In sectors like logistics or healthcare, AI-based scheduling systems have helped optimize resource use and respond more flexibly to demand fluctuations. Some workers report appreciating the removal of managerial

favouritism or inconsistency, particularly when algorithmic rules are clearly explained and contestable. Studies by the OECD (2023) suggest that algorithmic tools, when carefully designed, can support procedural fairness in routine decisions – such as leave approval or task rotation – so long as humans remain ultimately accountable. Still, these benefits are conditional: they depend not on the mere presence of AI, but on its alignment with organizational values, social dialogue, and legal safeguards.

AI and algorithms in the workplace

Under the AI Act, several uses of AI in the workplace are explicitly classified as high-risk:

- AI systems intended to be used to make decisions affecting the terms of work-related relationships, including hiring, promotion, task assignment, or contract termination;
- AI systems used to monitor and evaluate performance and behaviour of individuals in such relationships;
- AI systems that allocate tasks based on personal traits or characteristics, or that adapt workflows based on observed or predicted employee behaviour.

The particular practice of emotion recognition of employees is specifically banned under article 5 of the AI Act. The stated reason is the questionable nature of the underlying science in combination with the far-reaching impact this practice can have. The abovementioned high-risk practices similarly do have a far-reaching impact, but the associated compliance obligations (see chapter 3) do make the process transparent and subject to accountability (next chapter).

Specifically in the context of work many more regulations exist together with the AI Act. In many countries the works council has some form of right of consent or co-determination when it comes to employee evaluation systems, or at least must be consulted prior to its deployment. Under the GDPR, a fully automated employee evaluation system would almost certainly need a data protection impact assessment (DPIA) to mitigate any risks to the privacy or other fundamental rights of employees.

AI, algorithms and platform work

The Platform Work Directive (2024/2831, PWD), expected to be in force from 2026 onwards, establishes clear rights for workers subjected to algorithmic management in digital labour platforms. While originally targeted at ride-hailing, delivery, and microtask platforms, its requirements are increasingly being adopted as benchmarks in other AI-managed employment contexts, especially where work is mediated through apps or dashboards. These rules apply regardless of whether the worker is an employee or a freelance worker.

Under the PWD, platform workers must be informed when automated systems are used to make decisions that significantly affect their working conditions. This includes decisions related to task allocation, performance evaluation, pay calculation, or access to future work. Such systems must be transparent about their logic, significance, and expected consequences. In practice, this requires that workers receive intelligible explanations of how decisions are made, not just that automation is in use. In addition, significant decisions, such as deactivation, wage penalties, or contract non-renewal, must be subject to human review before taking effect.

Organizations that use algorithmic systems in platform-style environments should prepare governance mechanisms that include:

1. Worker access to plain-language documentation of system functionality;
2. Human review panels for contested decisions;
3. Internal risk assessments (e.g., FRIA or DPIA) specific to the employment context;
4. Joint consultation with worker representatives or unions on AI implementation and monitoring.

Employee de-skilling and up-skilling

ALTAI

F6. Could the AI system create the risk of de-skilling of the workforce?
F6a. Did you take measures to counteract de-skilling risks?
F7. Does the system promote or require new (digital) skills?
F7a. Did you provide training opportunities and materials for re- and up-skilling?

AI systems increasingly automate routine cognitive tasks – from customer support triage to document drafting and workflow decisions. While this can improve efficiency, it also raises a structural risk of de-skilling: the gradual erosion of human expertise, judgement, and situational awareness as workers come to rely on AI outputs rather than developing or exercising their own capabilities.

This phenomenon has been observed in sectors such as healthcare (with diagnostic support tools), finance (automated credit scoring), and logistics (predictive routing and scheduling). Over time, workers may begin to treat algorithmic recommendations as definitive rather than advisory, leading to a reduction in analytical vigilance or practical proficiency. In AI governance, this is referred to as **automation complacency**, and it directly undermines the goal of meaningful human oversight.

To counter this trend, organizations should implement safeguards that preserve and reinforce human skills. In particular, periodic re-training or simulation exercises are required to maintain decision-making capabilities independent of the system. More

generally, AI adoption generates demand for **new skills** – both technical and conceptual. Workers must learn to interact fluently with AI systems, understand their limitations, and identify when outputs require challenge or escalation. This requires targeted upskilling initiatives, particularly in:

- AI literacy: understanding basic system logic, confidence levels, and model types;
- Critical evaluation: spotting errors, bias, or inappropriate generalization in AI-generated outputs;
- Governance awareness: knowing when and how to raise concerns about system behaviour, bias, or harm.

AI in healthcare

In 2025, artificial intelligence is firmly embedded in the medical domain: from clinical decision support systems and diagnostic algorithms to triage automation tools. These systems influence high-stakes outcomes in a very visible way, and the risks associated with their deployment are qualitatively different from other sectors.

Clinical use of AI

AI systems are increasingly embedded in clinical workflows, most notably as tools for diagnostics, triage, image interpretation, and predictive decision support.[6] These systems are deployed in radiology, pathology, oncology, emergency medicine, and population health analytics, often assisting clinicians with tasks such as tumour classification, stroke detection, or predicting sepsis risk based on real-time vital signs. Their output generally influences clinical judgement, although final say remains firmly with the physician.

While clinical AI has advanced in accuracy and scope, it remains susceptible to critical errors. One notable risk is the generation of plausible but incorrect outputs, particularly from large general-purpose models that are adapted for medical use without domain-specific training or rigorous validation[7]. Such errors (commonly referred to as "hallucinations" in the broader AI literature) can lead to unsafe decisions when not flagged, reviewed, or contextualized by human clinicians. This is especially concerning in diagnostic support tools, where even a confident but flawed recommendation can unduly shape human judgement.

Moreover, some systems lack clear explainability, providing high-confidence assessments without transparent reasoning or clinical traceability. This can undermine both patient trust and professional accountability, particularly if the system's rationale cannot be reconstructed after a negative outcome. Clinicians risk being caught in a liability grey zone: responsible for decisions influenced by tools they may not fully understand or control.

To ensure safe and ethical integration of AI into healthcare, institutions must address:

1. Decision traceability: ensuring outputs are interpretable and auditable;
2. Clinical calibration: validating AI recommendations against gold-standard datasets and human benchmarks;
3. Defined human oversight: maintaining human-in-command protocols in all decisions that affect diagnosis, treatment, or care pathways.

Dual compliance: AI Act and Medical Device Regulation

Next to the AI Act, AI systems used in healthcare often face the stringent requirements of the EU's Medical Device Regulation (MDR). Generally, a medical device is any device – including software, such as an app or web service – that provides a specific medical purpose: diagnosis, prevention, monitoring, treatment of diseases or disabilities, investigations into the human body and examination of human specimens.

Like the AI Act, the MDR has its origins in product safety and similarly provides a clear path to compliance in the form of the conformity assessment process (discussed in chapter 3).[8] It recognizes four classes of medical device, from the low risk Class I to medium (IIa), higher (IIb) and high risk (III). Low risk is a simple non-invasive device, while high risk would be a device critical to sustaining life. Except for Class I devices, any AI system that serves as a safety component of a medical device, or functions as a standalone software medical device, qualifies as a high-risk AI system under Annex I the AI Act. (Additionally, the AI Act defines the practice of triaging for emergency healthcare as high-risk regardless of the MDR classification.)

The high-risk classification triggers the full set of AI Act obligations – data quality, transparency, human oversight, post-market monitoring, and conformity assessment – but does not replace MDR obligations; it adds a complementary layer focused on algorithmic behaviour and lifecycle governance. In practice, the MDR's conformity assessment would supersede the AI Act's similar process.

The CORE-MD clinical risk score was introduced in 2024 as a structured tool to assist regulators, manufacturers, and developers in evaluating the clinical risk level of AI-based medical device software.[9] Developed under the Coordinating Research and Evidence for Medical Devices (CORE-MD) consortium, the score considers:

- The system's intended medical purpose;
- Its level of automation or autonomy in clinical decision-making;
- The potential severity of harm if the system underperforms;
- The role of human oversight or dependency on AI output.

By using tools like the CORE-MD score and aligning internal development pipelines with both AI Act and MDR expectations, organizations can move toward regulatory maturity. In practice, this means:

- Starting MDR classification early in AI development;
- Designing audit trails and explainability features as part of AI Act compliance;
- Coordinating risk management and validation across both frameworks.

Ethical and professional reflections

From radiology and pathology to triage and mental health screening, AI systems are shaping how care is delivered, how risk is assessed, and how clinical decisions are supported. While these tools offer clear benefits in terms of speed, efficiency, and data processing, their integration into medical practice also raises questions about trust, autonomy, and the boundaries of professional judgement.

One of the most pressing concerns is the phenomenon of over-reliance. Clinicians may defer to AI-generated recommendations even when these contradict their own expertise.[10] A particular concern here is the time-pressing nature of decisions, which creates a self-fulfilling prophecy. Suppose an AI indicates a 70% risk of a patient needing to be intubated immediately to treat respiratory failure. Delaying intubation could increase the patient's mortality risk, which causes many physicians to opt for the recommended course, thus seemingly validating the AI's recommendation.[11]

Research shows that to achieve successful adoption, AI tools must: address real clinical pain points; demonstrate meaningful improvements in care; remain accurate and safe; integrate smoothly into existing workflows; and operate within transparent governance frameworks.[12] Clinicians are more likely to use AI tools when they understand how they work, when they help rather than hinder decision-making, and when they respect the relational and ethical fabric of medical practice. In the end, the goal is not merely to deploy AI – but to do so in a way that reinforces, rather than erodes, the practice of care.

Ethical concerns also extend to the long-term impact on clinical expertise. When key diagnostic or pattern recognition tasks are delegated to machines, clinicians risk a gradual erosion of skill. This mirrors the concept of de-skilling discussed in the previous section: physicians working alongside AI triage tools tend to exhibit lower diagnostic vigilance over time, especially in environments lacking routine feedback or manual review.[13]

These developments call for a broader reflection on the role of AI in the practice of medicine. The real transformation is not in speculative technologies like robotic caregivers, but in the quiet restructuring of clinical reasoning and workflow logic. AI is

changing not only how doctors make decisions, but how they experience responsibility, uncertainty, and trust. Preserving the human dimensions of care – judgement, empathy, and professional accountability– will be essential as AI becomes a permanent presence in modern medicine.

The limits of this human dimension become especially apparent now that patient-facing systems such as conversational agents, social robots and carebots are beginning to shape how individuals experience care. These systems are often deployed to support aging populations, assist with mental health monitoring, or provide companionship in long-term care settings. While they can offer practical benefits, such as reminders, monitoring, or even moments of comfort, they also raise serious ethical concerns. Multiple studies have noted that patients interacting with robotic or AI-based caregivers can experience a sense of emotional distance, artificiality, or even deception – particularly when the system mimics empathy without true understanding.[14] The risk is not that carebots fail to function – but that their use signals a broader shift toward depersonalized care, where the efficiency of the system is prioritized over the relational core of medicine.

AI in corporate sustainability and ESG governance

Artificial intelligence has become a material factor in corporate sustainability – both as a source of risk and as a tool for navigating complexity. What once fell under voluntary corporate social responsibility (CSR) now intersects with binding legal obligations under frameworks like the CSRD and CSDDD we discussed earlier. Organizations must account for the environmental and social impacts of their AI systems, from emissions and e-waste to human rights in algorithmically managed workforces. At the same time, AI technologies are being used to support ESG efforts themselves, helping companies analyze supply chain risks, forecast sustainability KPIs, and streamline disclosures.

From voluntary to mandated: ESG in the AI era

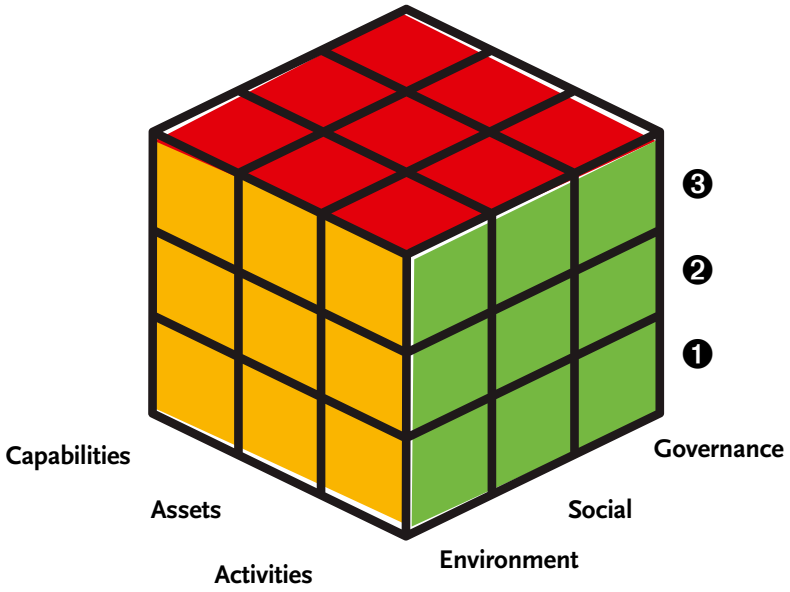
Environmental, Social, and Governance (ESG) concerns have evolved from aspirational commitments to enforceable obligations. The AI Act, CSRD, and CSDDD collectively mark the end of the era in which companies could treat sustainability and social impact as branding exercises. These frameworks demand systematic reporting, due diligence, and risk mitigation, from emissions and labor conditions to the development and deployment of AI systems.

For example, under the CSRD, companies must disclose how AI systems contribute to environmental impacts such as carbon emissions, water use, and electronic waste, as

well as social effects like algorithmic bias, worker surveillance, or automated decision-making in HR. The CSDDD further requires companies to identify, prevent, and mitigate adverse human rights and environmental impacts in their value chains – including those arising from outsourced AI development or platform-based labor models.

Operational tools for ESG

This shift from voluntary reporting to enforceable compliance creates both a burden and an opportunity. Organizations need clear, operational tools for embedding AI into ESG processes without reinventing internal governance. One such tool is Henrik Skaug Sætra's AI and ESG protocol, which forms a structured approach that connects traditional ESG dimensions with AI system design, deployment, and oversight.[15] As illustrated in the figure below, the protocol maps each ESG pillar to AI-specific concerns.



- **Environmental:** resource consumption (e.g. compute, electricity, water), emissions, e-waste
- **Social:** fairness, discrimination, labor impact, transparency of AI in customer interactions
- **Governance:** accountability structures, human oversight, ethical review, impact assessment

Sætra's framework is strategic overlay that helps companies align their AI practices with ESG goals. When paired with regulatory instruments like CSRD and the AI Act, it helps identify where AI needs to be surfaced in sustainability reporting, where due diligence is required, and where internal controls must be adapted to address new types of risk.

In practice, this means:

- Including AI-related indicators in ESG data pipelines (e.g. energy use per model, incidents flagged in algorithmic decision-making)
- Linking AI governance roles to existing ESG accountability structures
- Using impact assessments (see next chapter) to feed ESG disclosure and materiality analysis
- Embedding ESG and AI principles into procurement, especially for third-party models or platform services

By combining legal compliance with ethical structuring, organizations can move beyond baseline obligations and toward a credible, transparent approach to AI sustainability.

AI as a driver of ESG performance and reporting

While AI introduces sustainability and social risks that must be governed, it also plays an increasingly central role in delivering ESG objectives. Many organizations now rely on AI systems to process ESG-related disclosures, identify non-compliance in supply chains, monitor carbon footprints, and flag social risks such as labor exploitation or human rights violations. In this sense, AI is not only a subject of ESG governance, but also an operational enabler of it.

Under the CSRD, companies are required to disclose sustainability data across their operations and value chains. AI tools can support this by automating data extraction, generating metrics aligned with European Sustainability Reporting Standards (ESRS), and conducting natural language analysis on supplier documentation. Some ESG teams already deploy AI for:

- Screening for adverse media on supply chain partners
- Automating environmental impact assessments using satellite or sensor data
- Forecasting emissions trajectories based on production and logistics data
- Identifying greenwashing risks in communications or published disclosures

Yet AI's role in ESG reporting is not without caveats. Tools used for ESG must be explainable, auditable, and free from embedded bias. If, for example, an AI system scores suppliers for ESG risk but was trained on skewed or outdated data, the resulting decisions may reinforce existing inequalities. This makes model documentation, training data provenance, and periodic validation essential for credible ESG performance.

Organizations that rely on AI to deliver or support ESG efforts should:

- Conduct internal assurance of AI outputs used in sustainability reports
- Map AI-supported ESG indicators against ESRS data points
- Ensure human validation for critical classifications (e.g. "high-risk supplier" flags)
- Disclose, where relevant, the role of AI in generating or processing ESG data

Accountability and governance for AI-linked ESG risks

As AI becomes entwined with sustainability and social impact, organizations must ensure that ESG accountability frameworks are equipped to govern AI-related risks. This includes integrating AI into existing structures for risk oversight, materiality analysis, internal controls, and assurance. It also means clarifying who is responsible for managing the social and environmental consequences of AI use.

Under the CSRD, ESG disclosures must be signed off at the board level, and subjected to limited (and later reasonable) assurance. If AI systems contribute to Scope 3 emissions, labor impacts, or discriminatory outcomes, those effects must be surfaced in governance processes and reported transparently. The same applies under the CSDDD, where organizations can be held liable for failing to prevent adverse impacts—even when these occur via outsourced AI tools or services.

To meet these expectations, leading organizations are now adapting ESG governance models to include:

- **AI-specific risk registers:** documenting environmental, social, and human rights risks from AI systems, particularly those with public-facing, HR, or supply chain functions;
- **Expanded materiality processes:** evaluating whether AI systems materially affect ESG outcomes or stakeholder expectations;
- **Internal RACI matrices:** assigning responsibility for AI impact assessment, sustainability reporting integration, and incident escalation;
- **Cross-functional AI oversight committees:** embedding ESG, compliance, and tech experts into shared decision-making structures, often under the broader umbrella of AI governance (see Chapter 11).

This alignment of AI with ESG governance is not merely procedural. It reflects a shift in the nature of risk: from static to adaptive, from operational to reputational, and from siloed to systemic. The organizations best positioned for credible sustainability reporting re not those that "report" the most, but those that understand where technology intersects with environmental and social responsibilities – and manage it accordingly.

AI and democracy

AI LITAI

F8. Could the AI system have a negative impact on society at large or democracy?

AI systems are increasingly present in the institutions, processes, and communications that underpin democratic societies. From microtargeting in elections to algorithmic moderation of public discourse, AI is not just influencing how politics is conducted – but how it is perceived, shaped, and manipulated. This convergence raises serious questions: Could AI systems negatively impact democracy? Have their broader societal effects been assessed? And what safeguards are in place to minimize harm?

The democratic risk landscape

Today, AI technologies shape not only consumer markets and workplaces, but also the functioning of democratic societies. Algorithms determine what citizens see online, what narratives are amplified or suppressed, and which voices gain visibility in digital public spaces. From recommendation systems and political ad targeting to automated moderation and synthetic media generation, AI systems are deeply enmeshed in the infrastructure of public discourse and have direct consequences for pluralism, participation, and trust in democratic institutions.

A central concern is the algorithmic mediation of information. Recommender systems optimized for engagement can amplify outrage, reinforce echo chambers, or downrank minority or dissenting views. In doing so, they can distort the conditions for deliberative democracy, where informed, diverse perspectives are essential. Research from the past five years has shown that even slight changes in ranking logic or personalization parameters can tilt public debate, particularly when deployed by dominant platforms with vast reach.

Equally concerning is the rise of synthetic content—from deepfakes to AI-generated political commentary. These tools are increasingly easy to deploy, capable of mimicking human tone, style, and authority at scale. In electoral contexts, they can be used to impersonate candidates, sow confusion, or subtly manipulate public opinion through targeted misinformation. The risks are compounded by the velocity of digital media: once misinformation spreads, corrections rarely reach the same audience with equal force.

Finally, the opacity of AI systems in public administration—such as eligibility scoring, benefits allocation, or predictive policing—poses risks to transparency and institutional accountability. When citizens are affected by decisions they cannot trace or contest, public trust erodes, and democratic legitimacy is weakened. These risks

are not hypothetical: investigations across Europe have revealed multiple instances where opaque algorithms disproportionately affected marginalized groups or produced discriminatory outcomes.

The EU's 2022 Digital Services Act aims to regulate the large online platforms that today manage and control the online discourse. The issue of desinformation and algoritmic manipulation plays a central role in its application. One key weapon is its requirement of transparency: researchers must be able to do Big Data-style analysis on posts and engagements on these platforms. AI and machine learning algorithms may thus be able to learn to recognize fake news before it does harm. As noted in chapter 6 (transparency), generative AI system providers are required to label or watermark their outputs to make screening for it on social media platforms easier.

Assessing and minimizing societal impact

ALTAI

F8a. Did you assess the societal impact of the AI system's use beyond the (end-)user and subject, such as potentially indirectly affected stakeholders or society at large?

F8b. Did you take action to minimize potential societal harm of the AI system?

The societal effects of AI often extend far beyond the immediate user. A content moderation algorithm may influence what an individual sees, but its real impact lies in how it reshapes discourse across an entire platform. A voter-targeting tool may serve a single campaign, but its cumulative use across the electoral ecosystem can distort democratic participation, amplify disinformation, or erode public trust. These broader consequences are often diffuse, indirect, and cumulative – making them difficult to detect, but no less real.

To responsibly deploy AI systems with public or political relevance, organizations must systematically assess their second- and third-order impacts. This includes considering:

1. How system outputs might influence public discourse or shape narratives over time;
2. Whether the AI system systematically advantages or disadvantages certain groups;
3. Whether certain actors (e.g. governments, political parties, interest groups) might misuse the tool to manipulate public behavior or sentiment;
4. How unintended side effects might degrade civic trust, reduce pluralism, or exacerbate polarization.

Assessment is only the first step. Organizations must also take meaningful mitigating actions. These may include:

- Revising model goals or optimization parameters to reduce the amplification of polarizing content;

- Implementing guardrails such as caps on targeting granularity or disclosure overlays for synthetic content;
- Auditing training data and model behavior to detect unintended biases or socio-political asymmetries;
- Including civil society stakeholders in the design, testing, or impact review of systems used in sensitive civic contexts.

In politically charged or socially sensitive settings, mitigation may also require choosing not to deploy an AI system at all. The default assumption should not be that every application of AI is net positive – especially in contexts where harm to democratic institutions, minority rights, or civic trust cannot be confidently ruled out.

Safeguarding democratic integrity

ALTAI

F8c. Did you take measures that ensure that the AI system does not negatively impact democracy?

Preserving the integrity of democratic processes in the age of AI requires proactive restraint, design discipline, and institutional accountability.[16] While AI offers powerful tools for organizing civic participation, delivering public services, and analyzing social trends, it also opens new avenues for political manipulation, surveillance, and exclusion. These risks must be countered not just by individual developers, but by coordinated efforts across governments, platforms, and civil society.

The AI Act prohibits certain manipulative systems outright such as those exploiting vulnerabilities to distort behavior and classifies certain AI systems used in electoral processes, voter engagement, and political campaigning as high-risk. Other requirements can indirectly apply. For example, a chatbot interacting with voters during an election must be clearly identified as AI, its limitations disclosed, and escalation to a human made possible. Profiling systems used in campaigns must avoid infringing on rights to free expression or fair political competition.

Yet compliance is only the floor. True democratic safeguards require attention to design ethics, data provenance, and use context. Organizations should:

1. Disclose AI use in political communication, including microtargeting and content curation;
2. Maintain public registries of political AI systems, modeled after ad libraries or lobbying disclosures;
3. Limit personalization and predictive modeling in civic tools where it risks undermining voter autonomy or perpetuating exclusion;

4. Ensure human review and editorial accountability in AI-assisted journalism or content moderation tools;
5. Train operators and users of AI systems on civic responsibilities, freedom of expression norms, and risks of political bias or capture.

Platforms and institutions should also be prepared to suspend or decommission systems during electoral cycles or periods of political sensitivity if risks cannot be managed. Democracy is not an abstract principle – it is a system of accountability, transparency, and fair participation. AI systems that influence public reasoning, representation, or decision-making must be built to uphold those values by design, not merely through post hoc correction.

Key takeaways

Artificial intelligence is a systemic force with cross-cutting implications for society and the environment. From the energy use and water consumption of large-scale models to the acceleration of electronic waste, AI contributes to environmental burdens that must now be assessed and mitigated. In the workplace, AI reshapes task allocation, performance monitoring, and labor relations, raising concerns about transparency, de-skilling, and algorithmic oversight. In healthcare, AI tools support diagnosis and decision-making, but also challenge professional autonomy and patient trust, particularly when systems lack clarity, validation, or clinical accountability.

Across all these domains AI systems increasingly influence decisions with real-world social consequences. Organizations must move beyond narrow definitions of safety or performance to consider broader societal impacts: who benefits, who is excluded, and how core human values like fairness, dignity, and trust are preserved. Doing so requires a proactive commitment to ethical reflection, human oversight, and institutional transparency. These concerns converge in the next chapter, where we examine how accountability is structured, how harms are remedied, and how AI systems can remain answerable to the societies they shape.

10

Accountability and redress

Accountability is what makes an AI system reliable and trustworthy. Where transparency describes how a system operates, accountability demands to know why it was built that way, who is responsible for its outcomes, and what recourse exists when things go wrong. It is the structure that connects technical operations to ethical justifications and legal obligations. This chapter explores how AI systems can be made not only visible but answerable, and how organizations can ensure that their systems are not just understandable, but justifiable.

Understanding accountability

AI accountability, as outlined by the OECD, refers to "the expectation that organisations or individuals will ensure the proper functioning, throughout their lifecycle, of the AI systems that they design, develop, operate or deploy, in accordance with their roles and applicable regulatory frameworks, and for demonstrating this through their actions and decision-making process". [1] More generally, accountability mean the capacity to justify why a system behaves as it does, and who stands behind that behaviour. Transparency (see chapter 7) shows what happened; accountability explains why it was acceptable.

From traceability to justifiability

Traceability is a foundational requirement for AI accountability. It allows organizations to reconstruct how a system arrived at a particular output, what data it used, which model version was active, and how the decision path unfolded. This technical transparency is essential for audits, investigations, and debugging. But traceability alone does not make a system accountable.

To be truly accountable, an AI system must also be justifiable. That is, it must be possible to explain why the system was designed in a certain way, why particular decision logic was chosen, and why those choices are ethically and legally defensible. A logged recommendation is not self-justifying. A confusion matrix does not defend a hiring decision. These artifacts support retrospective analysis, but without an accompanying rationale and a responsible actor, they offer no meaningful accountability.

By the end of this chapter, you'll be able to ...

- Define key concepts underpinning AI accountability.
- Develop practical strategies ensuring AI accountability.
- Understand the relationship between AI accountability and societal trust.

Further, accountability is one step above compliance. Merely following the law to the letter is one thing; being accountable means being able to justify the choices made in the compliance process, or being required to redress what went wrong. This provides a three-step accountability process:[2]

- **Information**: The first step involves clearly outlining and providing knowledge about what the AI system does, its purpose, data it processes, algorithms it uses, and the decisions it makes.
- **Explanation or justification**: In the second step there's an onus on AI developers, managers, or deployers to offer reasons for why the AI behaves as it does. For instance, if an AI loan system denies an application, the system (or the organization deploying it) should be able to justify that decision. This is more than explaining *why* the decision was made (transparency, chapter 7) or ensuring the decision was fair (chapter 8). A justification means that the decision was right, legally and ethically – and with a clear up-front explanation rather than an after-the-fact excuse.
- **Consequences**: The third step relates to taking responsibility for AI outcomes. If an AI tool gives an incorrect or unjust output, there should be mechanisms for redress, which might include rectifying the mistake, compensating the affected person, or refining the AI model. There can be no accountability without consequences.

This distinction becomes urgent in contexts where AI systems make – or influence – decisions that affect people's rights, opportunities, or safety. In such cases, the justification must go beyond statistical performance. It must answer normative questions: Is the system fair? Were alternative designs considered? Who validated the deployment context? These are not questions of function, but of responsibility.

AI governance professionals must therefore treat traceability as a technical precondition and justifiability as an organizational and ethical obligation. One without the other leads to either obscurity or impunity. True accountability arises when both are present, and when the documentation that enables traceability is accompanied by clear reasoning, institutional ownership, and a willingness to respond to challenge.

Accountability as the normative core of trustworthy AI

Accountability is not just one pillar of trustworthy AI. It is the one that makes all the others enforceable. Fairness, robustness, and transparency are vital principles, but without accountability, they remain abstract ideals. It is accountability that translates these values into obligations: Who must ensure the system is fair? Who corrects it when it isn't? What happens if they fail?

The ALTAI framework defines accountability as the obligation to report, explain, and justify decisions made during the development and deployment of AI systems. It connects operational behavior to human responsibility, ensuring that AI does not float above the norms and structures of democratic societies. This view aligns with the OECD's definition, which emphasizes the need for clear roles, internal oversight, and redress mechanisms. Accountability, in this view, is both internal (within the organization) and external (toward affected parties and regulators).

In the literature, accountability has been described as a "relational obligation" that consists of three steps: the provision of information, the offering of justification, and the exposure to consequences.[3] This triad moves beyond documentation. It insists on meaningful engagement – on the ability of stakeholders to question decisions and demand change. It also distinguishes genuine accountability from its weaker forms: performative reporting, box-ticking compliance, or retrospective blame-shifting.

Role assignment and ownership

Accountability begins with clarity about who is responsible. No AI system can be held to account if the roles surrounding it are undefined, fragmented, or ambiguously distributed. Whether accountability is internal (for audit and escalation) or external (toward regulators or affected persons), it ultimately depends on identifiable agents – human or institutional – who can explain, justify, and, if needed, correct the system's operation.

Under the AI Act, this challenge is made explicit through legal roles: provider, deployer, importer, distributor, and authorized representative. Each has distinct obligations, as discussed in chapters 2 and 3. But in practice, these roles rarely map neatly onto organizational structures. A single company may fine-tune a model, integrate it with other components, host the service, and deploy it internally. This makes it both provider and deployer, creating potential liability in both capacities. In other cases, responsibilities are split across vendors, API providers, and internal product teams, raising the risk of diffusion or denial of accountability.

Establishing accountability, therefore, requires role mapping. This is the process of assigning responsibilities across the AI system lifecycle. This involves:

1. Identifying which actor determines the system's intended purpose;
2. Documenting who modifies models or logic;
3. Naming the party that interacts with affected individuals;
4. Clarifying which unit or person is responsible for oversight and escalation.

Effective role mapping is not only good practice – it is increasingly a compliance obligation. The AI Act requires providers and deployers to maintain documentation, designate responsibility, and ensure traceability. ISO/IEC 42001 reinforces this through the concept of accountable roles in an AI management system. And ALTAI asks directly: "Is it clear who is responsible for each step in the AI system's lifecycle?"

Designing for contestability

A system is not accountable if its actions cannot be challenged or contested. Contestability is the practical test of accountability: it asks whether affected individuals or oversight bodies can meaningfully dispute a system's decisions, require justification, and trigger a response. Most AI systems today are built for efficiency and optimization, not contestation. They aim to minimize latency, maximize throughput, and reduce human intervention. But contestability requires a different logic: systems must be designed not only to function, but to explain, to pause, and to revise. This means including mechanisms for:

1. Flagging outputs for human review;
2. Requesting the rationale behind a decision;
3. Registering and tracking complaints;
4. Revisiting decisions in light of new information or error discovery.

The AI Act reinforces this normatively by requiring that high-risk AI systems allow for human oversight, logging, and ex-post intervention. Without these capabilities, affected persons may be notified, but never heard.

Contestability must be built in before deployment. Retrofitting a complaints mechanism after harms have occurred will fail to restore trust or prevent recurrence. Moreover, designing for contestability supports better internal governance. When developers, risk officers, and compliance teams expect to be challenged, they build with clearer logic, better documentation, and greater humility.

Building for auditability and oversight

Accountability is a system of oversight, verification, and structured response. For AI systems, this begins with auditability: the capacity to trace how decisions are made, what data they rely on, and how they behave over time. Without auditability, accountability collapses into guesswork.

Traceability as foundation

ALTAI

G1. Did you establish mechanisms that facilitate the AI system's auditability (e.g. traceability of the development process, the sourcing of training data and the logging of the AI system's processes, outcomes, positive and negative impact)?

Auditability begins with traceability: the technical and procedural capacity to reconstruct how an AI system functions, and how it arrived at a given outcome. This means more than storing outputs. It requires versioning of training data, documentation of model iterations, and logs of runtime behaviour. Each of these elements allows organizations to track decisions back to their operational and developmental roots.

Modern AI toolchains increasingly support this traceability. Frameworks like DVC (Data Version Control), MLflow, and LakeFS allow teams to link models to specific datasets and parameter sets, creating a reproducible lineage of the training process. Similarly, structured logs of inputs, inferences, and outputs enable downstream actors to investigate anomalous behavior, understand edge cases, or validate claims of bias or error. These mechanisms are not optional: under the AI Act, they are core to both post-market monitoring and serious incident reporting.

But technical tools are only half the story. Traceability must be paired with governance processes that ensure logs are retained, accessible, and comprehensible. It must be clear who is responsible for maintaining traceability artifacts, how access is managed, and when audits are triggered. Systems developed without such foresight often lack the infrastructure to answer accountability questions after deployment—especially when ownership shifts between teams or external partners.

Enabling independent audit

ALTAI

G2. Did you ensure that the AI system can be audited by independent third parties?

True accountability requires that an AI system be auditable not only by its creators, but by third parties, such as regulators, internal oversight bodies, external experts, or affected stakeholders. Establishing this kind of openness demands deliberate design. It starts with transparency of the development process: documenting key design choices, training methodologies, risk assessments, and validation procedures. These records must be structured in a way that is accessible to non-developers, and without requiring source code review or machine learning expertise.

Second, auditability depends on technical modularity. Systems with tightly coupled components or opaque dependencies hinder external review. By contrast, AI systems

that isolate core inference modules, annotate their inputs and outputs, and maintain interface logs enable auditors to assess outcomes without needing to reverse-engineer the full stack. This aligns with guidance emphasizing auditability through deliberate design choices and reproducible processes across the AI lifecycle.[4]

Black-box models are not inherently un-auditable, but systems that entangle components (e.g. data pipelines, inference engines, interface logic) without clear boundaries create audit friction. Isolating decision points, annotating input/output behavior, and logging intermediate steps all support meaningful third-party review – whether for internal audit, regulatory inspection, or public accountability.

Third, organizations must address governance readiness. Independent audit is not just about exposing the system; it's about structuring access and accountability. This means having:

- Designated contacts for audit engagement;
- Access protocols that protect sensitive data while enabling transparency;
- Clear documentation of roles and responsibilities across the lifecycle;
- A process to act on audit findings, including escalation and remediation.

As the OECD has noted, when third parties can verify that a system's claims are substantiated by evidence, it strengthens both regulatory credibility and public legitimacy.[5] This is especially true when audits extend beyond development into deployment and monitoring phases, where real-world drift and unanticipated outcomes often emerge.

A growing body of implementation-focused work shows how this can be done in practice. For example, Helmer et al. demonstrate how audit requirements can be translated into concrete tasks within the MLOps lifecycle, using decision records and traceability artifacts that serve both internal quality assurance and external audit readiness.[6] By integrating audit checkpoints across stages of development and aligning them with project maturity levels, organizations reduce friction at the point of external review.

Ethics committees, audits, and oversight mechanisms

ALTAI

G3. Did you foresee any kind of external guidance or third-party auditing processes to oversee ethical concerns and accountability measures?

G3a. Does the involvement of these third parties go beyond the development phase?

Establishing oversight mechanisms is one of the most direct ways to institutionalize accountability. While audit trails and documentation create the technical infrastructure, oversight bodies and ethical review processes provide the organizational

scaffolding for evaluating and contesting AI system decisions before, during, and after deployment. A 2023 literature review identifies key risk drivers such as unrepresentative data, cognitive bias in system interpretation, and trade-offs between efficiency and due diligence, and concludes that periodic audits, human oversight, and governance structures are essential to ensuring AI systems remain fair, accountable, and socially beneficial over time.[7]

One effective model is the creation of cross-functional AI ethics committees. These bodies bring together technical leads, legal experts, compliance officers, and external stakeholders to review use cases, flag grey zones, and deliberate on trade-offs in system design. Such committees are particularly valuable in contexts where legal rules offer little guidance on ambiguous harms or emerging risks, as will be the case with most new AI deployments.

In parallel, external auditing of ethical accountability has become a structured offering within the professional services sector. Major professional consultancy firms now routinely offer recurring ethical assurance for AI systems, especially those operating in regulated or rights-sensitive environments. These audits go beyond traditional IT compliance and assess whether a system's design, deployment, and monitoring practices reflect principles such as fairness, explainability, and accountability over time. Importantly, this work is increasingly embedded into lifecycle management.

These structures are not new: healthcare ethics boards, research review panels, and data access committees have long played similar roles. The novelty lies in adapting them for algorithmic systems, where the complexity of models, the scale of deployment, and the opacity of outcomes can obscure ethical harms. A formalized review process forces teams to articulate system objectives, defend risk mitigation strategies, and justify normative assumptions. This creates a documented space for ethical reasoning, not just technical compliance.

Risk awareness and organisational learning

ALTAI

G4. Did you organise risk training and, if so, does this also inform about the potential legal framework applicable to the AI system?

The most robust logging system or external audit will fail if the people involved in designing, deploying, or overseeing an AI system cannot recognize risk, interpret ethical tension, or respond to emerging concerns. Accountability requires that individuals within an organization understand not just what the rules are, but why they matter, and how they apply to real-world decisions.

Training beyond compliance

Training is often treated as a compliance checkbox: a brief module on acceptable use, a PowerPoint about AI risks, a signed policy at onboarding. But these approaches fall short. Effective training must go beyond awareness and build the ability to recognize, assess, and act on risk across technical, legal, and ethical domains.

AI systems introduce specific challenges that require tailored capacity-building. Developers must understand the implications of model design choices, not only in terms of performance, but fairness and downstream impact. Risk managers must be able to distinguish between statistical uncertainty and normative risk. Product owners must learn when a feature crosses into regulatory territory. Without a shared language and conceptual grounding, accountability becomes fragmented.

Training must therefore be role-specific and context-sensitive. Developers need guidance on dataset bias, explainability trade-offs, and documentation standards. Compliance officers need fluency in the AI Act's obligations, GDPR interfaces, and audit expectations. Operational users must understand what outputs mean, when to escalate concerns, and how to interact with fallback mechanisms. A one-size-fits-all training program cannot deliver this.

Equally important is ongoing reinforcement. A one-time training at system launch quickly fades in relevance. Organizations must create routines, such as refresher sessions, post-incident reviews and embedded ethics prompts, that support continuous learning. The erosion of human judgment is a key risk in algorithmic systems, and regular training is a key countermeasure.[8] When people become disengaged or defer to the system, accountability fails.

Framing and addressing ethical ambiguity

ALTAI

G5. Did you consider establishing an AI ethics review board or a similar mechanism to discuss the overall accountability and ethics practices, including potential unclear grey areas?

Many accountability failures in AI do not result from negligence or bad intent, but from ethical ambiguity—those "grey zones" where the law is silent, where values come into tension, or where trade-offs are poorly understood. These scenarios are not rare edge cases. They are the everyday reality of deploying AI systems in messy, contested human environments. And they cannot be solved by rules alone.

Organizations must learn to deliberate in uncertainty, using structured processes to surface ethical questions early, discuss them openly, and escalate them when needed. These

discussions may not always yield a clear answer, but they prevent silent normalization of harm. Without them, questionable design choices slip through development unnoticed, only to resurface as regulatory violations or reputational damage.

One useful approach is to classify and prepare for recurring ethical ambiguities using an internal Ethical Grey Zone Matrix. This is a practical tool to help teams identify, categorize, and respond to ambiguous ethical situations that arise during AI development and deployment. Rather than waiting for external scrutiny or internal

Grey Zone Type	Description	Questions to Ask	Escalation Strategy
Value Conflict	Tension between competing principles (e.g. accuracy vs. fairness)	Which value has priority in this context? What precedent exists?	Ethics committee or cross-functional review
Legal Underspecification	No clear regulation applies or law is outdated	What is the regulatory intent? What guidance do similar domains provide?	Legal counsel + compliance escalation
Group-Level Impact	Effects are diffuse, systemic, or not tied to individuals	Who is affected indirectly? Are structural inequalities being amplified?	Societal impact assessment or stakeholder input
Proxy Risk	A variable is legally permitted but may serve as a proxy for protected traits	Could this data encode bias? Has it been tested across demographic groups?	Bias testing + data governance review
Silent Drift	A system shifts over time, creating misalignment with original ethical goals	Is the current behavior still aligned with intent? Are new risks emerging?	Post-market monitoring + trigger review

conflict, this matrix encourages proactive reflection. By framing typical grey zones (typically value conflicts, proxy risks or system drift) and pairing them with guiding questions and escalation strategies, organizations can build shared awareness and a consistent approach to navigating ethical uncertainty.[9]

Such tools create a shared vocabulary for discussion. When teams can name a risk as a proxy issue or value conflict, they can better collaborate on solutions. Moreover, structured deliberation promotes transparency in decision-making: not just what was done, but why.

In practice, this requires safe spaces and structured time to raise ethical concerns. Teams under pressure to ship features rarely prioritize reflection. Embedding lightweight review points at sprint retrospectives, architecture review boards, or project milestones can normalize this work. Over time, the organization builds ethical reflexes: the ability to sense ambiguity and respond with clarity, not silence.

The role of internal expertise and organizational memory

As systems evolve, are updated, and passed between teams, continuity in ethical reasoning and institutional responsibility can be easily lost. This is where organizational memory becomes essential: the ability to retain, retrieve, and act on prior decisions, justifications, and lessons learned.

Organizations often focus on audit logs and documentation for traceability, but these artifacts are only meaningful when contextualized by people with historical understanding. Who approved a system for deployment? What trade-offs were considered during a redesign? Which ethical concerns were raised, and how were they resolved? Without answers to these questions, the same issues are likely to reoccur, and sometimes with more serious consequences.

To maintain this memory, organizations must invest in internal expertise—not just technical leads or compliance officers, but individuals or teams tasked with curating and communicating the governance history of AI systems.[10] This may include:

- Maintaining decision records and ethical review summaries;
- Running post-incident reviews that feed into training and documentation;
- apturing “soft signals” from prior deliberations that never made it into formal policies;
- Supporting knowledge transfer when staff leave or teams change.

Internal expertise also enables reflexivity. By fostering a culture that values reflection, organizations can gradually build ethical maturity. Accountability then becomes more than a safeguard; it becomes a source of strategic strength, as teams learn to anticipate ethical risk and respond with informed, confident decisions.

Continuous ethical alignment

ALTAI

G6. Did you establish a process to discuss and continuously monitor and assess the AI system's adherence to the ALTAI Assessment List?

G6a. Does this process include identification and documentation of conflicts between the 6 aforementioned requirements or between different ethical principles and explanation of the 'trade-off' decisions made?

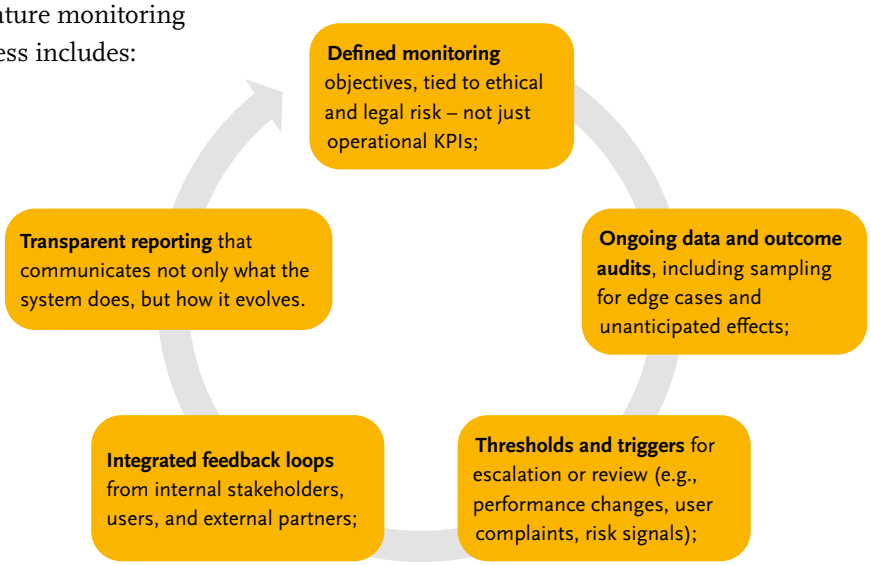
G6b. Did you provide appropriate training to those involved in such a process anddoes this also cover the legal framework applicable to the AI system?

AI systems evolve over time: models are updated, data shifts, contexts change, and risks emerge in ways that were not visible at the point of deployment. A system that was once compliant and justified may, over time, drift into ethical misalignment. For this reason, accountability must include mechanisms for ongoing monitoring, reassessment, and adjustment.

Monitoring legal and ethical adherence over time

Once a system is live, its performance, impact, and alignment with ethical and legal standards must be monitored continuously. This is a governance necessity in fast-changing social and technical environments. Effective monitoring requires that organizations move beyond narrow performance metrics. While accuracy and error rates are essential, ethical adherence involves additional dimensions: fairness drift across user groups, emerging misuse patterns, breakdowns in human oversight, and deviations from the system's originally justified purpose. These issues often develop gradually and may not trigger alarms unless systematically tracked.

A mature monitoring process includes:



This work is continuous and should be structured as part of the organization's quality management or risk governance functions.

Handling ethical trade-offs and value conflicts

AI systems often create situations where two legitimate values come into tension: fairness and accuracy, privacy and personalization, transparency and security. These are not bugs; they are structural dilemmas. Yet without a process to handle them, such dilemmas are either ignored or resolved informally, leaving organizations vulnerable to both ethical drift and public backlash.

To be accountable, organizations must explicitly address ethical trade-offs and value conflicts as part of their governance process. This means creating structured forums where such conflicts can be surfaced, deliberated, and documented.

A practical framework for handling these conflicts includes:

- **Identification**: Clarify which values are in tension and where in the system they arise (e.g., model design, data choice, interface decisions).
- **Impact assessment**: Analyze how each path affects different stakeholders, especially those historically marginalized or affected by automation bias.
- **Precedent review**: Examine how similar conflicts have been handled in the organization or sector to ensure consistency.
- **Justification**: Clearly articulate the rationale behind the decision, including the principles prioritized and the mitigations proposed for the deprioritized value.
- **Documentation**: Log the trade-off decision in governance records to support future review, transparency, and accountability.

Mäntymäki refers to this as translating ethical principles into "actionable governance routines."[10] Without this translation, even the best-articulated values remain abstract, and decision-making defaults to technical feasibility or business expediency.

Documenting trade-offs does not guarantee correctness. But it ensures that decisions are made consciously, with traceable reasoning, rather than as the incidental outcome of unexamined bias or system constraints. It also creates organizational memory: when the same trade-off arises later, teams have a starting point for discussion and improvement.

Equipping ethics functions with legal capacity

Ethics teams and committees are often tasked with navigating complex value conflicts and high-stakes decisions. Yet many of these bodies operate without a clear understanding of the evolving legal landscape. What's more, too often, ethics reviews are treated as advisory: their recommendations can be overruled, delayed, or ignored in the rush to deploy.[4] What's missing is not only a voice in the room, but the procedural weight to affect outcomes and the legal knowledge to know when intervention is not optional but required. Thus, ethics functions must be equipped with legal literacy and procedural authority.

To bridge this gap, organizations should integrate legal capacity into ethics governance through:

- **Cross-training**: Ethics teams must be conversant in relevant regulatory frameworks and obligations across jurisdictions. This doesn't require full legal expertise, but a working understanding of what triggers compliance duties or legal risk.
- **Legal liaison roles**: Embedding or assigning legal counsel to ethics committees ensures timely input on regulatory implications of technical or design decisions.
- **Shared knowledge bases**: Creating accessible repositories that link ethical principles to relevant legal standards (e.g., transparency under AI Act Art. 13 vs. GDPR Art. 15) supports consistent interpretation.
- **Scenario-based learning**: Ethics teams should periodically review real or simulated cases where legal obligations and ethical tensions intersect – such as profiling without consent, explainability gaps, or silent model drift.

This fusion of ethics and legal awareness also protects against the misperception that ethics is "soft" while law is "hard." In reality, many ethical concerns signal legal risk – especially where AI impacts fundamental rights, discrimination, or due process. Anticipating these risks early, and understanding how legal frameworks apply, strengthens both ethical decision-making and legal defensibility.

External reporting and redress by design

What distinguishes accountable organizations is not perfection, but preparedness: a willingness to listen, investigate, and respond. Accountability must include mechanisms that allow external stakeholders to report issues and seek redress. This is especially critical in high-impact domains where individuals may be subject to automated decisions without clear paths for objection or correction.

Risk and bias reporting channels

ALTAI

G7. Did you establish a process for third parties (e.g. suppliers, end-users, subjects, distributors/vendors or workers) to report potential vulnerabilities, risks or biases inthe AI system?

No accountability framework is complete without a way for people outside the system's development to signal when something is going wrong. Internal testing, oversight, and monitoring can miss important risks. For instance, risks may only emerge in small subpopulations or occur under unusual real-world conditions. That's why organizations must establish clear, accessible channels for reporting vulnerabilities, bias, or harm from external parties: users, workers, suppliers, and even bystanders.

These channels must go beyond generic contact forms or privacy disclaimers. They need to be designed with intent. This includes:

- Dedicated escalation routes for AI-related concerns, distinct from general product or customer support lines;
- Named points of contact or functional roles (e.g. ethics liaison, risk officer);
- Anonymity and non-retaliation guarantees, especially when employees or partners report risks;
- Accessible interfaces and language for affected users, including non-technical audiences;
- Clear framing of what types of issues can be reported – bias, unfair outcomes, explainability gaps, misuse, etc.

Importantly, these mechanisms must work in both directions: enabling individuals to raise concerns, and ensuring that organizations treat those concerns seriously. This means implementing triage protocols, assigning reports to responsible staff, and defining timelines for investigation, response, and potential corrective action. Otherwise, reporting becomes performative only.

Establishing and advertising these channels also supports regulatory compliance. Under the AI Act, high-risk systems must be supported by a Post-Market Monitoring (PMM) system, as detailed in Chapter 3. With this structured process real-world performance data is collected in order to identify and mitigate emerging risks. Receiving feedback directly from affected persons is a valuable way to acquire information on risks that own channels would not be able to. Even outside of formal compliance, organizations that treat external voices as part of their governance structure are better positioned to build long-term trust.

Feedback loops and corrective governance

ALTAI

G7a. Does this process foster revision of the risk management process?

Reporting channels are only the beginning. What defines an accountable organization is how it responds when concerns are raised. For AI systems, especially those deployed in sensitive contexts, feedback must be more than acknowledged – it must be processed, escalated, and converted into improvement. This is where feedback loops come in: internal governance processes that translate signals into action.

Unfortunately, many organizations still treat incident response as an informal task, dependent on individuals rather than process. Reports are received but not routed. Issues are discussed but not resolved. Corrections, if made, are undocumented. Over time, these gaps accumulate – not as technical debt, but as governance drift: the erosion of institutional capacity to identify, own, and fix ethical problems.

To counter this, organizations must implement structured feedback cycles that span technical, legal, and operational domains. Below is a comparative illustration of the difference between ad hoc response and mature corrective governance:

Governance Function	Ad Hoc Response	Structured Feedback Loop
Triage	Issues handled case-by-case, based on individual discretion	Reports classified and routed using defined severity categories
Ownership	No clear point of contact for follow-up	Accountability assigned to named roles or functions
Investigation	Informal inquiry, often disconnected from dev teams	Linked to model documentation, audit logs, and risk registers
Correction	Fixes applied ad hoc, rarely documented	Actions tracked in PMM system and governance logs
Learning	Lessons undocumented; issues reoccur	Post-incident review feeds into design, policy, and training

Under the PMM process, feedback from external reporting must be linked to investigation and remediation. Without this connection, even well-intentioned risk channels become dead ends. Organizations using MLOps workflows are best positioned to support this kind of lifecycle accountability.[6] Their integration of audit catalogs into iterative development allowed complaints or ethical concerns to trigger code-level reviews, architectural adjustments, and even rollback protocols.

Contestability and redress in practice

ALTAI

G8. For applications that can adversely affect individuals, have redress by design mechanisms been put in place?

The final test of accountability is whether individuals affected by an AI system can meaningfully challenge it. This is the principle of contestability, the right to object, to receive justification, and to seek redress. For many AI systems, especially those influencing decisions in credit, employment, healthcare, or public administration, contestability is not just an ethical ideal; it is a legal requirement.

To make contestability real, organizations must design for redress. This means structuring systems and processes so that individuals can:

1. Recognize that an AI-driven decision has been made;
2. Understand its consequences and the basis for its outcome;
3. Access a channel to dispute or appeal the decision;
4. Receive a timely response with appropriate remedies where applicable.

Organizations often assume redress means financial compensation or a formal apology. But redress, in the AI context, must be more tailored to the nature of the harm. If a candidate is excluded from a job shortlist due to a biased screening model, the appropriate redress is not money, but a re-review of the applicant's file, this time without the tainted filtering logic. If a credit limit was reduced based on a faulty risk model, redress may include not just reinstatement, but a manual override process and a transparent explanation. In automated public services, redress might involve restoring lost access (e.g., to benefits), removing harmful records, or issuing a formal correction to a person's profile. In each case, redress is about restoring a fair process, not just compensating for a bad outcome.

Some harms may not be individual at all. A flawed AI deployment might produce group-level harm, for instance by disproportionately rejecting applications from a particular postcode, ethnicity, or age bracket. Here, redress could involve identifying and notifying affected individuals, offering them a renewed opportunity to engage with the system

under fair conditions and taking additional steps to restore any lost access, status, or opportunity they were unfairly denied. This is distinct from correcting the system itself; redress is about making people whole.[11]

But building for contestability is more than a service design problem. AI systems must be configured to support reversibility. This includes maintaining decision logs that allow reviewers to reconstruct what data was used, which rules were applied, and whether human oversight was properly exercised. It also means empowering humans in the loop—not just to observe the system, but to overrule it when needed.

Reflecting with ALTAI: Ethics in practice

In the compliance-heavy context of the AI Act, it is easy to lose sight of the fact that trustworthy AI also requires deliberation. That is where the ALTAI framework comes in. Developed by the European Commission's High-Level Expert Group on AI, ALTAI (Assessment List for Trustworthy AI) is not a legal instrument or risk classifier. It is a tool for ethical reflection, designed to surface tensions, trade-offs, and areas for improvement in the design and deployment of AI systems.

Deploying ALTAI in an organisation

While ALTAI offers a solid foundation for assessing the trustworthiness of AI systems, its real-world value lies in how it is adapted and integrated into an organisation's specific context.[12] As such, it should evolve alongside the technologies and use cases it is meant to assess. Thus, it must be tailored to reflect the organisation's goals, values, and operational realities. This includes modifying the question set, assigning weights based on risk relevance, or adding custom criteria. For example, in high-stakes domains like autonomous vehicles, researchers have successfully sector-adapted ALTAI to address domain-specific challenges.[13]

For organisations with existing compliance frameworks (e.g. GDPR, ISO/IEC 27001), ALTAI should be seen as a complementary ethical layer, not a competing process. Integration works best when done deliberately and with broad stakeholder support.
A streamlined integration approach involves:

1. **Gap analysis** of current frameworks;
2. **Stakeholder alignment** across legal, technical, and governance teams;
3. **Customization** of the ALTAI questions and scoring;
4. **Training** for teams involved in AI development and review;
5. **Feedback and iteration**, to avoid box-ticking and ensure relevance;
6. **Periodic updates**, aligned with changes in systems or regulations.

Organisations aiming for structured maturity under ISO/IEC 42001 will find ALTAI a useful starting point for building AI governance awareness, especially in early-stage or ethics-by-design workflows.[14]

The spider chart: Visualizing trustworthiness

ALTAI's most recognizable feature is the spider chart (sometimes called radar chart) – a visual summary of how an AI system scores across the seven dimensions of trustworthy AI. Rather than delivering a score or certificate, the spider chart highlights strengths, weaknesses, and blind spots in a way that facilitates discussion. It helps teams ask: *Where are we strong? Where are we overlooking impacts? Where do we need more reflection – or intervention?*

Each axis of the spider chart corresponds to one of ALTAI's seven topics. To populate the chart, organizations assess their AI system using ALTAI's self-assessment questions, assigning scores to each question or sub-topic. While ALTAI does not prescribe a fixed scoring method, a common approach is to use a weighted point system:

- Each question can be assigned a value (e.g. 1–5 points), depending on its relevance and criticality to the system in question.
- Some lower-impact questions may be scored as 1 point ("yes/no" or basic checks), while others that probe substantive issues – like bias mitigation or human oversight escalation paths – may be worth 3 to 5 points.
- Organizations may also choose to **omit questions** that are not applicable or adjust their weighting to reflect internal priorities or sectoral risks.
- It is encouraged to **add organization-specific questions** where ALTAI's baseline does not fully address unique ethical concerns, such as public communication in controversial deployments or accessibility for vulnerable user groups.

The total score per axis is then normalized or scaled to fit the spider chart. The result is not a compliance verdict, but a **reflection profile** – a way to visualize the ethical shape of the system, spot gaps or imbalances, and guide informed discussion among stakeholders.

Examples of ALTAI in action

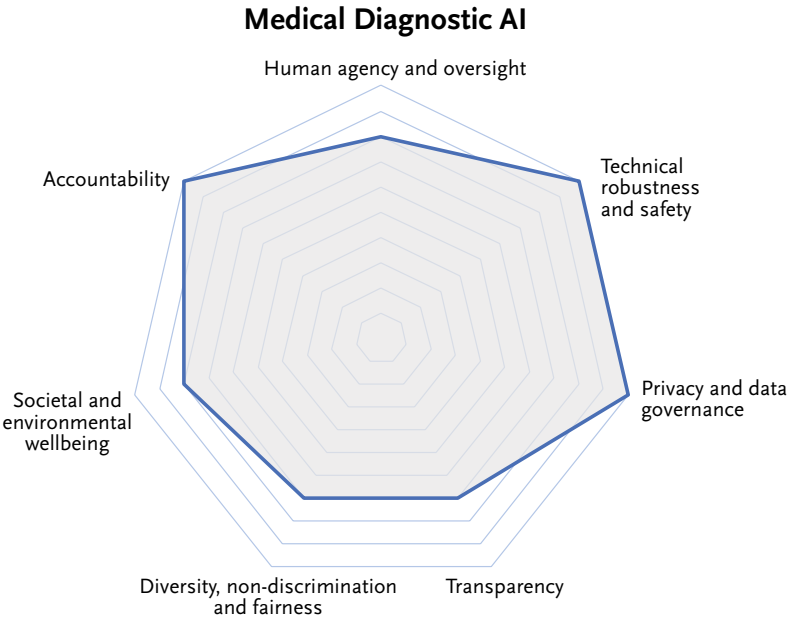
To show how context shapes trustworthiness, we present three example cases, each with a distinct spider chart:

Medical Diagnostic AI

This system analyzes medical data such as images, lab results, and patient histories to diagnose diseases or conditions. It aids healthcare professionals by providing insights based on patterns that might be challenging for the human eye to detect, ensuring early and accurate diagnosis.

- **High scores** in robustness, technical safety, and accountability, reflecting regulatory pressure and strong documentation.
- **Lower scores** in transparency and societal impact, where user communication and unintended consequences (e.g. over-reliance) were underdeveloped.

Insight: The spider chart helped the hospital identify a need to invest in user explainability training and fallback procedures in case of low-confidence outputs.



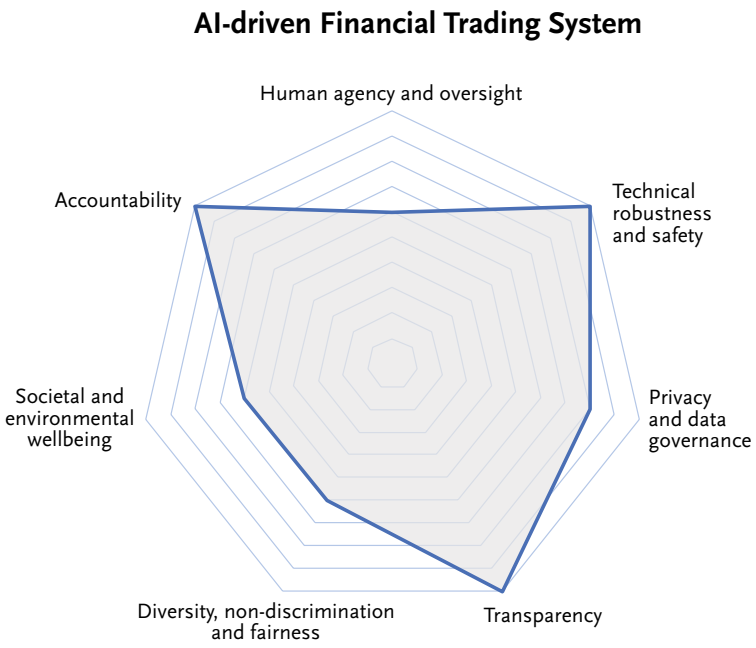
AI-Driven Financial Trading System

Utilizing real-time market data, historical trends, and complex algorithms, this system predicts market movements and executes trades at optimal times. It's designed to maximize profits while minimizing risks, operating at a speed and precision beyond human capabilities.

- **Strong in performance and autonomy**, with high marks for technical design and minimal manual intervention.

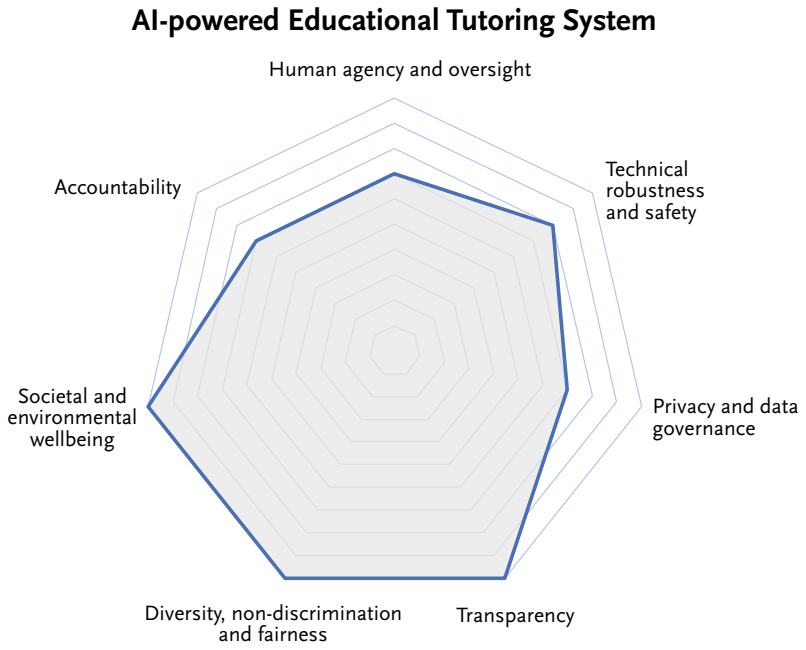
- **Weak in fairness and transparency**, as the model is opaque even to its developers and its external impacts (e.g. market volatility) were not assessed.

Insight: The ALTAI chart made visible the system's ethical asymmetry: it was optimized for internal efficiency, but blind to societal effects or contestability.



AI-Powered Educational Tutoring System

Tailored to individual student needs, this system offers personalized learning experiences. It assesses students' strengths and weaknesses, adapts content accordingly, and provides real-time feedback, ensuring a more effective and engaging learning process.



- **Moderate scores across the board**, with strengths in transparency and human oversight.
- **Lower scores** in fairness and data governance, especially regarding representation in training sets and the system's effects on student autonomy.

Insight: The spider chart triggered further review by the school's ethics committee, which recommended periodic re-evaluation of datasets and additional oversight when used with underperforming students.

Key takeaways

Accountability is the connective tissue that links AI performance to human responsibility, oversight to redress, and governance to trust. We have seen how accountability is built into AI systems through auditability, risk ownership, ethical review, and feedback loops. But accountability must extend outward, allowing affected individuals and communities to raise concerns, contest outcomes, and receive meaningful redress. Importantly, redress is not a synonym for apology or compensation. Rather, it is the process of restoring fairness, re-opening opportunity, and making harm visible and repairable.

But these mechanisms do not operate in isolation. They require organizational structures that are capable of learning, adjusting, and holding themselves to account. This brings us to the final layer of AI governance: the institutional practices, roles, and cultures that support long-term accountability.

AI Governance in the organisation

Managing AI means governing a powerful, evolving system that cuts across every layer of the organization. *AI governance* refers to the strategic and operational frameworks that ensure AI systems are developed and used responsibly, safely, and in alignment with both legal obligations and ethical values. Unlike compliance, which ensures adherence to specific rules, or risk management, which focuses on avoiding harm, governance encompasses the full lifecycle of AI systems and aligns AI initiatives with organizational purpose, stakeholder trust, and societal expectations. Organizations must not only comply with requirements but also demonstrate internal capacity to govern AI responsibly.

What is AI governance, and why does it matter?

Organizations are increasingly expected not only to comply with legal requirements but to actively govern how AI is developed, deployed, and monitored. Governance is what turns AI from a technical asset into a managed organizational function. It ensures that the use of AI aligns with institutional goals, public expectations, and legal obligations.

Defining AI governance

AI governance refers to the institutional frameworks, processes, and responsibilities that oversee the design, development, deployment, and monitoring of AI systems.[1] It is not limited to meeting external rules or preventing harms, but instead encompasses the broader task of steering AI technologies in line with organizational purpose, legal mandates, and ethical expectations. Governance involves setting internal policies, assigning roles, establishing review procedures, and ensuring ongoing accountability for outcomes produced by AI systems.

By the end of this chapter, you'll be able to ...

- Understand the comprehensive landscape of AI ethics and law.
- Navigate compliance using the ALTAI assessment tools and implement best practices.
- Stay adaptive and responsible in the ever-evolving AI domain.

It is important to distinguish governance from compliance and risk management. Compliance is about adhering to specific rules, such as the transparency obligations or data quality standards outlined in the AI Act. Risk management, in turn, focuses on identifying and mitigating potential harms, whether technical, legal, or reputational. Governance, by contrast, provides the strategic structure within which both compliance and risk are managed. It integrates these functions with decision-making authority, operational procedures, and value-driven oversight. In short, governance is not just about doing things right – it's about deciding what should be done in the first place, and how responsibility for it is organized.

A challenge for AI governance is the simple question of what exactly are we governing when we talk about "AI"? Artificial Intelligence is not a singular technology, but a diverse collection of applications enabled by the global system of distributed computing.[2] From this perspective, AI is not a discrete invention but a general functionality of the digital ecosystem, rooted in the interplay of data, computing power, networks, and software. This ecosystemic framing helps explain why the governance challenges of AI long predate recent developments like large language models. The real governance problem is not a particular technique, but the scale, reach, and institutional embedding of automated decision systems across domains. Attempting to govern "AI" in the abstract would imply asserting control over the entirety of distributed computing. A more effective and proportional approach is to govern specific applications of machine learning, where impacts, risks, and responsibilities can be meaningfully defined and allocated.

Governance as a lifecycle function

AI governance is not a fixed task to be completed at deployment, nor a reactive exercise triggered by regulatory scrutiny. It is a continuous function that must accompany an AI system across its entire lifecycle. Each phase in the lifecycle introduces distinct questions of responsibility, oversight, and documentation, as set out in the table overleaf:

In practice, this means that organizations must establish governance mechanisms that are embedded into the operational reality of AI development. For example, a use case intake process should include not just technical feasibility but also an assessment of regulatory exposure, rights impact, and alignment with organizational values. During system design and development, governance structures should ensure that data sourcing, model validation, and documentation meet applicable standards and can withstand future audit. Once deployed, systems must be monitored for degradation, unintended outcomes, or shifts in AI Act risk classification (see chapter 3).

Lifecycle Phase	Key Governance Activities	Typical Roles Involved
❶ **Intake**	• Use case assessment • Initial risk and impact screening • Regulatory classification	Use Case Owner, Legal, Compliance
❷ **Design & Development**	• Data sourcing approvals • Model documentation • Bias checks and reproducibility logs	ML Engineers, Data Stewards, Compliance
❸ **Testing & Validation**	• Evaluation against performance and fairness criteria • Internal review & sign-off	Technical Leads, Ethics Board, Product Manager
❹ **Deployment**	• Logging & traceability setup • User documentation • Role-based access control	IT/Operations, Security, Data Protection Officer
❺ **Monitoring & Oversight**	• Drift detection • Performance auditing • Reclassification if context changes	Compliance, Risk Management, Supervisory Authorities (if needed)
❻ **Decommissioning / Update**	• Retirement planning • Sunsetting schedule • Lessons-learned integration	Product Owner, Governance Lead, Records Management

Governance as an organizational responsibility

AI governance is an organization-wide responsibility that emerges from how roles, decisions, and oversight mechanisms are distributed across teams. While compliance obligations may formally fall on a provider or deployer, effective governance depends on coordinated action across legal, technical, operational, and strategic functions. This is particularly true under the AI Act, which places concrete obligations on both providers and deployers. Fulfilling these requirements cannot be outsourced to a single compliance officer. It demands an integrated governance structure capable of linking strategic intent to operational execution.

Organizations must therefore translate governance into clearly defined responsibilities, escalation pathways, and accountability chains. This includes assigning owners to AI use cases, embedding compliance checks in design workflows, and maintaining internal registries of deployed systems. Governance becomes real not through abstract principles, but through everyday operational decisions: Who signs off on a system's deployment? Who monitors model drift? Who is responsible when something goes wrong? A 2025 literature review confirms that most AI governance research lacks real implementation focus, with only a handful of studies addressing the full scope of governance – who is responsible, what should be governed, when interventions are needed, and how they are enacted.[3] This reinforces the need for governance approaches that are not only principled, but embedded in practice, driven by clear roles, lifecycle processes, and integration across legal, ethical, and operational domains.

Governance roles and responsibilities

AI governance only works when responsibility is made real: assigned, understood, and embedded in daily operations. The complexity of AI systems, and the breadth of legal and ethical requirements attached to them, demand more than ad hoc committees or post-hoc sign-offs. Governance must be operationalized through clearly defined roles across the AI lifecycle, from intake to monitoring.

Core roles in AI governance

Effective AI governance depends on people. Within an organization, AI governance must be carried by a network of roles that span legal, technical, ethical, and operational domains. These roles are not abstract titles but embedded functions, each with specific responsibilities in assessing, shaping, and overseeing AI systems as they move through the organizational lifecycle.

At the center of this network is the AI Compliance Officer (CO), an emerging role designed to anchor AI governance in both legal rigor and ethical reflection. The CO is not simply a rule enforcer; their strength lies in navigating complexity, identifying systemic risks, interpreting regulatory frameworks across jurisdictions, and translating those into actionable guidance for teams. Their training emphasizes not just knowledge of the AI Act and related legislation, but also the ability to work autonomously, think critically, and steer organizational decision-making with sensitivity to context and resistance. In practice, the CO operates as a connector between domains: reviewing documentation with Legal, challenging modeling decisions with Data Science, raising human rights concerns with the Ethics Board, and ensuring ongoing alignment with the organization's risk appetite and compliance posture.

Alongside the compliance officer, domain-specific roles play equally vital parts. Risk Officers or enterprise risk management functions ensure that AI systems are evaluated not only for legal or reputational exposure but also in terms of broader operational, financial, and strategic risk. This includes evaluating model robustness, business continuity dependencies, and vendor-related exposures for off-the-shelf or API-based AI services. Legal Counsel brings expertise in contractual liability, intellectual property, and regulatory interpretation, while the Data Protection Officer (DPO) safeguards compliance with data protection law—ensuring that personal data used in training, inference, or logging meets GDPR standards, and that data subject rights are respected across the lifecycle.

Technical roles such as ML Engineers, Data Stewards, and Product Owners are equally embedded in governance, not only through execution but through accountability. They are expected to make governance "real" by ensuring systems are traceable, explainable, and auditable in practice – not just in principle. Importantly, these teams must be trained to recognize when governance issues arise and to escalate appropriately, whether to Legal, Risk, or Ethics. A well-functioning oversight structure does not rely on central gatekeepers but enables distributed responsibility, supported by training, workflows, and clarity about when to escalate and to whom.

The role of the Ethics Board or Oversight Committee

In organizations that take AI governance seriously, ethical oversight is institutionalized through designated review bodies. These may be called ethics boards, AI oversight committees, or responsible innovation panels, but they share a common purpose: to provide structured, multidisciplinary judgment on the social, legal, and ethical implications of AI initiatives.[4] Their role is advisory, but should come with real influence.

An effective ethics board does not duplicate compliance reviews but complements them. It focuses on the grey zones: where the law is silent, where risk is uncertain, or where organizational values might be compromised by speed, ambition, or commercial pressure. In such cases, the board functions as a deliberative space for challenge and course correction. Its composition is critical – typically involving members from legal, technical, operational, and external domains (e.g., civil society or academic ethics). This diversity ensures that decision-making is not captured by a single perspective and that trade-offs are surfaced rather than buried. (Compare the substantive discussion in the previous chapter on continuous ethical alignment.)

The review literature[3] notes that very few governance frameworks provide a concrete structure for ethical oversight. Most remain focused on high-level principles or procedural risk management, missing the critical role that institutional deliberation

plays in trust-building and conflict resolution. This underscores the value of embedding ethical review into the governance workflow, not as a hurdle but as a structured pause where the organization reflects on the downstream consequences of AI deployment. Whether integrated into project approval gates or convened for escalated cases, the ethics board is a central mechanism for translating principles into decisions.

Mapping responsibilities with a RACI Matrix

One of the most persistent failures in AI governance is ambiguity: everyone is responsible, yet no one is accountable. To avoid this, organizations must clarify *who does what* across the AI system lifecycle. The RACI model—short for Responsible, Accountable, Consulted, Informed—is a proven tool for mapping out decision rights and governance duties.[5] Applied to AI, it helps align technical, legal, ethical, and managerial roles across key stages of system development and use.

The RACI matrix distinguishes four types of involvement:

- **Responsible**: the role(s) doing the work.
- **Accountable**: the ultimate owner of the outcome; only one per task.
- **Consulted**: those providing input or review.
- **Informed**: those who need to be kept updated.

Below is a simplified example illustrating how typical roles map to stages of AI governance. Note that these are illustrative; organizations must adapt based on structure, risk exposure, and regulatory role (e.g., provider vs. deployer).

The strength of this model is not in the boxes, but rather in the discussions it forces. Who should really be accountable for risk classification? When should the ethics board be consulted, and when does it need to be informed? Organizations that answer these questions early build resilience and reduce the likelihood of governance-by-surprise.

Frameworks and maturity models

Frameworks and maturity models provide the scaffolding that allows organizations to embed governance practices in a repeatable, scalable, and auditable way. But with the proliferation of models, ethical guidelines, risk management tools, compliance checklists, and management system standards, organizations often face a new challenge: not the absence of frameworks, but the difficulty of selecting and aligning them.

Lifecycle Stage	AI CO	Legal	DPO	Risk Officer	ML Engineer	Product Owner	Ethics Board
Use Case Intake	C	C	I	R	I	A	C
Risk Classification	A	C	C	R	I	R	C
Design & Development	C	I	C	I	R	A	I
Deployment Approval	A	C	C	I	R	A	C
Monitoring & Escalation	A	C	C	R	R	I	C

Governance frameworks in use: Ethics, Risk, and Compliance

AI governance frameworks serve different functions depending on what they are designed to manage. Some aim to surface ethical reflection, others to ensure legal defensibility, and still others to embed continuous improvement in system design and use. Understanding the distinctions and relations between these frameworks is essential for building a coherent governance strategy.

At one end of the spectrum lies ALTAI (Assessment List for Trustworthy AI), the self-assessment tool that serves as the foundation of this book. It is not a legal instrument, nor is it designed to support regulatory conformity. Its function is primarily ethical: to help organizations think through issues such as fairness, human agency, and societal impact. ALTAI is ideal for early-stage discussions, ethics board deliberations, or organizations developing internal codes of practice. However, it is not sufficient as a standalone compliance tool (although it can be turned into one).

Further along the spectrum are impact assessment frameworks, notably the AI Act's Fundamental Rights Impact Assessment (FRIA) that we discussed last chapter. FRIA, DPIA and other such frameworks serve a distinct governance function: they structure the identification, classification, and documentation of risks. Unlike ethical reflection tools, which surface abstract tensions or value conflicts, risk-oriented assessments require more systematic inquiry: What could go wrong? Who is affected? Under what conditions does the risk materialize? These frameworks act as a bridge between early-

ALTAI
- Promote internal reflection on trustworthy AI principles
- Voluntary
- Ethics Boards, Compliance Officer, Strategy Leads

FRIA
- Identify and document impacts on fundamental rights, data protection, and legal risk
- Mandatory (certain high-risk deployers)
- Legal, DPO, Risk Officers, Compliance Officer

CA
- Demonstrate system compliance with AI Act
- Mandatory for high-risk AI
- Providers, Legal, Technical Documentation Teams

stage ethical awareness and formal governance mechanisms by enabling organizations to capture and manage risk before systems are too far along to meaningfully intervene.

At the compliance end of the spectrum are conformity assessment procedures mandated under the AI Act. These formal processes require organizations (especially providers) to demonstrate that their systems meet legal requirements for data governance, transparency, human oversight, and more. The exact procedure varies by risk tier and system type (e.g. internal checks vs. third-party assessment), but the outcome is binding: legal authorization to place or use a system on the EU market.

Standards and systemic models: ISO/IEC 42001, NIST AI RMF

While tools like ALTAI and FRIA help organizations reason about individual AI use cases, systemic governance models are designed to govern across portfolios. These are not checklists or point-in-time assessments—they are management systems and frameworks that embed AI governance into organizational structure, culture, and process. Two of the most prominent are ISO/IEC 42001 and the NIST AI Risk Management Framework (AI RMF).

ISO/IEC 42001 is the world's first formal AI Management System Standard (AIMS).[6] Published in late 2023, it offers a structured approach for organizations to design, implement, maintain, and continuously improve an AI governance system. Like its ISO predecessors (e.g., 27001 for information security), 42001 introduces a Plan-Do-Check-Act cycle and requires documented policies, roles, monitoring mechanisms, and internal audits. It is certifiable, making it suitable for organizations that want to demonstrate maturity and trustworthiness to external stakeholders.

The NIST AI RMF, developed by the U.S. National Institute of Standards and Technology, is non-certifiable but highly practical.[7] It is structured around four core functions – Map, Measure, Manage, and Govern – and is designed to help organizations identify and mitigate risks in AI development and use. What sets the AI RMF apart is its accessibility: it is modular, sector-neutral, and supported by a growing body of implementation guidance. Many organizations, especially outside the EU, use it as a governance entry point or as a flexible complement to ISO standards.

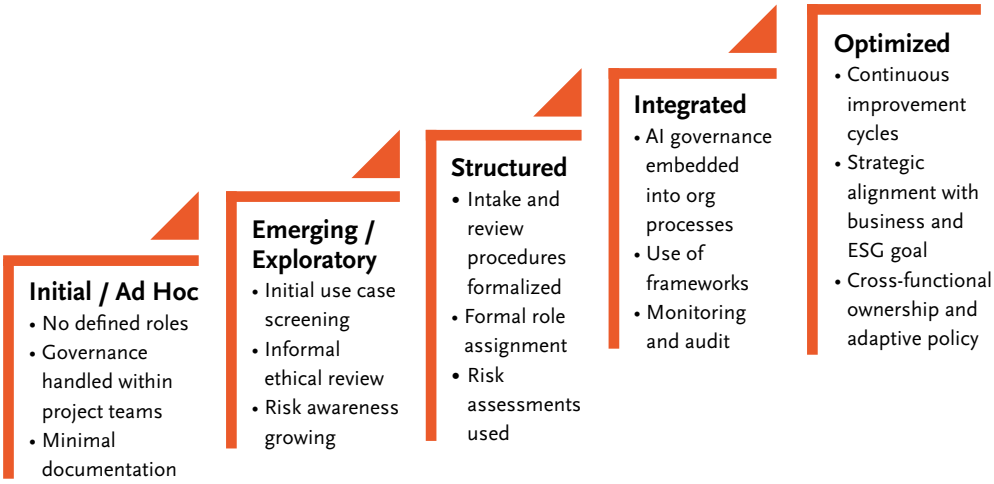
The AI Act does not prescribe any particular standard for AI governance. Rather, it requires an adequate risk management and quality management systems to be in place. In this context, frameworks like ISO/IEC 42001 and the NIST AI RMF offer structured and credible approaches for building those systems. ISO 42001, in particular, provides a certifiable foundation that organizations can adopt to formalize their internal governance architecture, while the NIST framework offers a flexible, principle-based model to identify and manage AI-specific risks.

What these frameworks do not provide is a so-called presumption of conformity (see chapter 3). That is only available after compliance with EU-approved harmonised standards. Applying other frameworks may implement conformity, but requires additional proof to market surveillance authorities that these frameworks are sufficient.

Governance maturity models and gap assessment

Governing AI is a progressive capability that must grow with the complexity and scale of AI adoption. Understanding this growth becomes possible with the AI Capability Maturity Model (AICMM) developed by Hansen et al.[8] Drawing from case studies of organizations at different stages of AI implementation, the model presents governance maturity as a function of organizational diffusion.

The AICMM identifies how governance tasks evolve as organizations progress from ad hoc experimentation to systemic deployment. At lower levels of maturity, governance is informal and reactive: decisions are made within project teams, documentation is sparse, and oversight is episodic. As maturity increases, organizations begin to institutionalize practices by introducing structured intake, defined accountability, risk tracking, and internal review mechanisms. Eventually, AI governance becomes embedded in enterprise-wide management systems, aligned with corporate strategy, and subject to continuous improvement.



What distinguishes AICMM from checklist-based approaches is that it recognizes governance as a learning function. It accounts for the fact that early-stage teams may require flexibility and speed, while later stages demand consistency, traceability, and strategic alignment. This makes it especially useful for guiding not only current-state assessments, but also roadmaps for capability development. Rather than forcing all teams to adopt mature processes prematurely, the model supports proportionality.

This dynamic, developmental view of governance is echoed in system-level frameworks like ISO/IEC 42001 and the NIST AI RMF, both of which encourage regular review, stakeholder engagement, and documentation of improvement actions. But AICMM provides the missing layer: a model grounded in how real organizations grow. As noted earlier, many governance efforts fail because they focus narrowly (typically on compliance, ethics, or risk) without a strategy for coherence or progression.[3] Maturity models fill that gap.[9]

Practical governance workflows

AI governance becomes real through workflows: the structured, repeatable processes by which organizations review, approve, monitor, and retire AI systems. These workflows connect high-level governance goals such as fairness, accountability and legal compliance to day-to-day actions across teams. Without such operational structures, even the best intentions remain theoretical. Applying the AICMM from the previous section, let's examine what such workflows would look like.

Use case intake and review

Every AI governance process begins or fails at intake. The use case intake stage is where the organization decides whether a proposed AI system is appropriate, permissible, and governable. Without a structured intake process, high-risk systems may enter development pipelines without proper scrutiny, leading to costly redesigns or non-compliance down the line.

A mature intake workflow begins with a use case proposal, typically initiated by a product owner, business lead, or technical team. This proposal should include not only the functional goal of the system, but also its operational context, target users, data dependencies, and expected level of autonomy. The use case card discussed in chapter 3 can be a useful instrument here.At this stage, the goal is not to approve or reject, but to classify the system according to its risk profile, impact domain, and applicable regulatory requirements.

The AI compliance officer (CO) plays a central role here. Working alongside legal counsel, risk officers, and data protection teams, the CO ensures that the intake process includes an initial assessment of risk level (e.g. per AI Act categories), relevance of existing governance frameworks (e.g. need for FRIA, DPIA), and any triggers for ethical review. Where applicable, the intake may also route the proposal to an internal ethics board or oversight committee for deliberation, especially if the system involves human rights implications, vulnerable populations, or opaque decision logic.

A well-designed intake process does more than gatekeep. It creates a traceable record of how and why an AI system entered the development lifecycle, what initial risks were identified, and who was consulted. This documentation becomes the foundation for downstream activities such as impact assessment, conformity preparation, and eventual audit. It also reinforces governance as a collaborative process, not an afterthought.

Monitoring and risk reclassification

In chapter 3 we discussed the AI Act's legal obligation of post-market monitoring (PMM), the structured process of monitoring deployment of high-risk AI systems with the aim of spotting risks early, preferably before actual damages occur. The way such monitoring is implemented reveals whether the organization can detect when an AI system begins to exhibit risk and whether it has the processes, roles, and authority in place to respond. This makes PMM a governance benchmark.

Following the AICMM, monitoring is not treated as a compliance function alone, but as a governance capability. Early-stage organizations may perform monitoring sporadically or rely on developers to flag issues informally. As maturity increases,

monitoring becomes a formalized activity with assigned responsibility, clear intervals, and integration into risk management processes. Importantly, it is coupled with the authority to act, notably by reviewing decisions, updating impact assessments, or triggering reclassification. Effective risk management is not simply about identifying risks but about sustaining the organizational capacity to act on them in structured, documented, and context-sensitive ways.[10]

Risk reclassification is a key function of this governance maturity. AI systems have a tendency to evolve, often through expanded use, new data inputs, or changes in downstream decision-making. A system that begins in a low-risk context may accumulate risk over time. Governance workflows must therefore support the possibility of moving a system into a higher oversight tier, and doing so without friction or delay. This depends on having escalation mechanisms that are not just reactive but understood, rehearsed, and institutionally owned.

For AI compliance officers, the governance question is not "are we monitoring?" but "can we respond?" Monitoring only supports responsible AI if it feeds into a larger architecture of oversight, escalation, and system-level adaptability. In this way, PMM becomes a living signal of how well governance is functioning, not simply whether the rules are followed, but whether the organization is capable of recognizing when they need to be updated or enforced more stringently.

Incident escalation and oversight

Even in the most mature AI governance environments, systems will fail, behave unexpectedly, or raise questions that no predefined policy can fully answer. What distinguishes a resilient organization is not the absence of incidents, but the presence of clear, practiced pathways for escalation and oversight. This is where governance moves from documented intent to organizational reflex.

An incident in the context of AI governance is any event that signals a deviation from expected system behavior, risk boundaries, or legal/ethical norms. This may include legal issues (e.g. a discriminatory output or data protection breach) but also business or technical issues: a spike in user complaints, unintended use cases, or system decisions that pass technical checks but raise ethical concerns. What matters is not only identifying the incident, but activating the right people, at the right time, with the right authority to intervene.

A mature escalation process is built around three elements: classification, routing, and resolution authority. First, incidents need to be categorized: Is this a technical error, a policy breach, or a novel situation requiring ethical judgment? Second, they must be routed to

the correct responsible entity (see under governance roles above). Finally, the organization must have clarity on who decides: Can the system be paused or rolled back? Must it be reported to regulators? Does it trigger a formal review or a revision to governance policies?

Organizations operating at higher maturity levels (as described in AICMM) tend to treat escalation not as an admission of failure, but as a designed function of governance. Escalation logs are maintained, patterns are analyzed, and outcomes are fed back into training, policy updates, and system design reviews. The incident response process becomes not just a safeguard, but a source of institutional learning.

Sunsetting and system decommissioning

Most AI governance discussions focus on how systems are built, validated, and monitored. Far fewer address how they end. Yet the responsible retirement of AI systems is as critical to governance as their deployment. Systems may become obsolete, unmaintainable, misaligned with business goals, or out of step with evolving legal, social, or organizational standards. Without a defined process for sunsetting, outdated systems may linger in production, generating silent risks and accountability gaps.

In governance terms, sunsetting is a lifecycle event, not a failure. Mature organizations include decommissioning criteria in the initial intake and risk classification process, identifying triggers such as declining accuracy, end of business relevance, unresolved audit flags, or replacement by newer models. From the AICMM perspective, the ability to retire systems in a controlled, auditable, and transparent way signals high governance maturity. It reflects not just technical discipline but institutional foresight.

A structured decommissioning workflow should include:

- A decision protocol: Who decides that a system should be retired, and on what basis?
- A transition plan: How will dependencies, user impact, and business continuity be managed?
- A data governance review: What happens to logs, models, training data, and audit trails?
- A closure record: What lessons were learned, and how are they fed back into future governance?

Without a decommissioning process, governance becomes sticky: systems accumulate, responsibilities blur, and accountability fades. In contrast, organizations with clear sunsetting procedures demonstrate that AI governance is truly lifecycle governance. They don't just control what enters the system; they manage what stays, and they know when to let go.

Governance and compliance integration

As organizations mature, the challenge is not simply to govern AI well, but to integrate AI governance into existing systems of control, from data protection and cybersecurity to risk management, internal audit, and sustainability. Building trusted AI requires a deliberate effort to identify shared controls and align objectives across business units and governance domains.[11] AI systems may introduce new risks, but they also intersect with long-standing frameworks for information security, financial risk, ESG compliance, and operational integrity. Governance, in this context, becomes a cross-functional architecture: one that scales, coordinates, and evolves with the enterprise.

Mapping AI governance onto internal control architectures

As organizations adopt AI, one of the most overlooked governance challenges is not how to create new controls, but how to reuse and align existing ones. Effective AI governance begins with recognizing how it fits into the organization's broader Governance, Risk, and Compliance (GRC) infrastructure. The task of governance, then, is to identify which functions need extension, which controls can be inherited, and where new responsibilities must be defined.

Everything starts with control mapping. Many of the mechanisms required for responsible AI – access management, incident escalation, data lineage tracking, auditability – already exist in other parts of the organization. Information security teams enforce access logs and anomaly detection; compliance units oversee third-party vendor controls; risk officers manage registers of critical exposures. AI governance does not need to rebuild these; it needs to connect them. A well-integrated governance program builds bridges, not silos.

For example, the AI compliance officer might work with IT risk to extend existing asset inventories to include AI systems. Instead of creating a separate risk taxonomy, AI risks can be mapped into existing enterprise risk categories (e.g. operational, reputational, legal). Audit procedures used for financial systems can inform the traceability expectations for high-risk AI models. The result is not just efficiency—it's consistency: a unified governance posture that enables auditability, reduces duplication, and avoids "compliance fatigue" across teams.

Integration with GDPR, cybersecurity, and ESG programs

As AI systems increasingly influence decisions about individuals, process sensitive data, and impact social or environmental outcomes, aligning AI governance with existing programs (notably GDPR, cybersecurity, and ESG/CSRD) is both practical and necessary.

Take data protection. The GDPR already requires Data Protection Impact Assessments (DPIAs) for systems that pose high risks to individual rights and freedoms. High-risk AI systems, especially those using personal data for profiling, must therefore be assessed through both an AI-specific lens (e.g. FRIA, risk classification under the AI Act) and a data protection lens. Rather than treat these as parallel efforts, mature organizations synchronize DPIA and AI governance workflows, using shared templates, role structures (e.g. the DPO), and escalation channels.

The same principle applies to cybersecurity. Frameworks like ISO/IEC 27001 and NIST CSF already include controls for system availability, data integrity, and incident response. High-risk AI systems similarly need such controls, especially when deployed in critical infrastructure or embedded in real-time decision-making. AI governance should not duplicate cybersecurity functions, but rather extend them: for instance, ensuring that model monitoring includes not only accuracy drift but exposure to adversarial manipulation or system compromise.

Finally, AI systems increasingly factor into environmental, social, and governance (ESG) reporting, particularly under the Corporate Sustainability Reporting Directive (CSRD) and related European Sustainability Reporting Standards (ESRS), discussed in chapter 9. Organizations are expected to report on social impacts such as discrimination risk, accessibility, and the effects of algorithmic decision-making on stakeholders. AI governance provides the documentation and oversight mechanisms to support these disclosures.

In all three domains the same lesson applies: AI is not an isolated object of compliance, but a horizontal risk that intersects with vertical compliance domains. The role of governance is to build structured bridges between them, ensuring that AI systems are not only lawful and ethical, but also integrated into the organization's existing fabric of control.

The Anti-Silo principle: Governance as organizational glue

One of the most persistent risks in AI governance is fragmentation. As organizations build up parallel efforts in data governance, cybersecurity, legal compliance, and ethical oversight, AI often becomes the common thread—but is treated as a separate initiative. The result is duplication of controls, inconsistent decision-making, and missed opportunities for shared learning. Mature governance does the opposite: it acts as organizational glue, connecting these domains through shared language, workflows, and accountability.

This is sometimes called the *anti-silo principle,* the idea that AI governance should not introduce a new layer of bureaucracy, but rather serve as a coordination function across existing structures.[12] Instead of creating new committees or documentation pathways, governance teams should plug into existing forums (e.g. risk committees, product approval boards, IT change control) and align AI oversight with already institutionalized review cycles.

This anti-silo approach does not mean reducing ambition. On the contrary, it allows AI governance to scale. When intake forms mirror existing project management templates, when risk classifications align with enterprise risk taxonomies, and when monitoring uses shared infrastructure, governance becomes faster, more intelligible, and harder to ignore. It embeds itself into how the organization already operates, rather than demanding a parallel world.

Most importantly, anti-silo governance fosters shared ownership. AI compliance officers, legal teams, data scientists, and product leads no longer act in isolation or pass responsibility between departments. They engage in joint decision-making, grounded in a common understanding of roles and thresholds. This is not just efficient—it is what makes governance credible.

Oversight, audits, and governance maturity

As AI systems grow in complexity and risk, the ability to show that governance processes exist, are followed, and are improving becomes a strategic necessity. This applies not only to regulators and auditors, but to boards, partners, and the public. Oversight and auditability are not final steps: they are integral to building trust, refining governance processes, and signaling organizational maturity.

Internal oversight structures

Before regulators, auditors, or external stakeholders ask questions about how AI is governed, the organization must be able to answer those questions itself. Internal oversight structures provide the mechanisms through which AI governance becomes visible, contested, and improved.[13] They turn policy into practice and ensure that governance is not confined to documents—but enacted through deliberation, challenge, and coordination.

A mature internal oversight structure is not a single body or committee – it is a network of functions working together. This includes AI compliance officers, legal and risk leads, IT architects, product managers, and internal audit functions. Together, they form a distributed governance ecosystem, each responsible for specific stages in the

AI lifecycle, with clearly defined escalation pathways and joint review responsibilities. What matters is not that every issue is solved by a central committee, but that there is clarity on who sees what, when, and what actions follow.

Some organizations establish dedicated AI governance committees; standing bodies that review high-risk use cases, track emerging risks, and align governance decisions with strategic goals. Others embed AI oversight into existing governance forums such as risk committees, innovation boards, or data ethics panels. The key is integration. AI oversight must be part of the organization's broader control framework, rather than an isolated structure disconnected from where actual decisions are made.

The AI compliance officer plays a critical role here, not just as a gatekeeper but as a coordinator. They ensure that oversight is consistent, that policies are followed, and that deviations are escalated or addressed. They also help structure documentation, facilitate cross-functional reviews, and maintain institutional memory across use cases.

Audit readiness and external demonstrability

Audit readiness is the operational capacity to show, with evidence, that policies are in place, processes are followed, and risks are being managed.[14] It is what turns governance from an internal ideal into a verifiable reality for external stakeholders: regulators, certification bodies, customers, or civil society.

Whether through internal checks or third-party evaluation, organizations must provide structured documentation: technical files, impact assessments, risk logs, human oversight measures, and post-market monitoring protocols. Audit readiness is not about scrambling to meet these requirements at the end but about building them into the lifecycle from intake to decommissioning. (Compare last chapter's section on building for auditability for practical tips.)

From a governance perspective, audit readiness requires two core capabilities: traceability and retrievability.[15] Traceability means that key decisions about the AI system (notably, design choices, risk classification and review outcomes) are recorded and linked to responsible roles. Retrievability means that this information is stored in a structured way that can be surfaced quickly during internal reviews or external audits. These requirements extend beyond technical teams: they involve legal, compliance, risk, and operational staff working with a shared understanding of what must be documented, how it is maintained, and who is accountable for its accuracy.

Using maturity models for continuous improvement

Governance evolves with the organization's systems, risks, and institutional learning. To support this evolution, maturity models offer a structured way to assess where you are, where gaps exist, and what capabilities must be built to govern AI systems more effectively.[16] This book builds on the AI Capability Maturity Model (AICMM) by Hansen et al., which positions AI governance as a function of how deeply governance practices are integrated into operational, ethical, and strategic layers of the organization.[8]

The AICMM tracks the transition from ad hoc responses to structured and reflexive governance systems, emphasizing that oversight must grow in parallel with AI adoption and diffusion. What makes the model actionable is not just its categories, but the steps it suggests to move from one stage to the next. Below, we adapt the AICMM into a practical roadmap focused specifically on governance maturity, with concrete actions organizations can take at each stage.

Level	Governance Posture	Practical Actions to Advance
❶ **Ad Hoc**	No formal oversight or documentation	• Identify all current AI use cases • Assign initial ownership (e.g. product owner, CO) • Conduct a baseline risk scan (basic classification)
❷ **Emerging**	Governance exists in pockets; awareness rising	• Develop a simple AI use case intake form • Introduce case-by-case review process • Assign CO to coordinate early-stage assessment • Start logging decisions informally
❸ **Structured**	Defined governance roles and processes	• Formalize intake and review workflow • Establish oversight board or integrate into existing forums • Require FRIA/DPIA for high-risk systems • Maintain a system registry with versioning, documentation, and contacts
❹ **Integrated**	Governance embedded in core business and tech operations	• Align AI governance with internal audit and risk functions • Connect to data protection and cybersecurity programs • Set up monitoring + reclassification triggers- Review all active systems on a scheduled basis
❺ **Optimized**	Continuous improvement based on feedback and metrics	• Use incident logs and audit outcomes to revise policy • Define and track KPIs (e.g. time to review, number of escalations) • Run governance retrospectives after major deployments • Benchmark against external standards or peer organizations

This structure gives governance teams something they can act on: not a theoretical score, but a list of operational upgrades to work through. It also avoids the pitfall of over-building—by encouraging just enough structure for each stage of maturity.

External oversight and supervisory engagement

As AI systems become more embedded in public services, financial decisions, and critical infrastructure, external stakeholders (regulators, journalists, works councils, civil society groups) are playing a growing role in shaping expectations and demanding accountability. Mature governance anticipates this. It is not only inward-facing but externally legible, ready to engage with oversight mechanisms beyond the organization.

From a governance perspective, external oversight demands three capabilities:

1. **Proactive transparency** – Organizations should prepare "governance briefing kits" that describe their risk classification approach, escalation procedures, and monitoring mechanisms in clear, structured language. These materials should be accessible not only to technical peers but to legal, supervisory, and public audiences.
2. **Regulatory fluency** – AI governance teams, especially the AI compliance officer, must understand the requirements of the AI Act and other applicable laws (e.g., GDPR, sectoral regulation). They must be able to map internal processes to legal expectations and engage constructively with regulators during audits, inquiries, or consultations.
3. **Structured engagement** – Beyond compliance, organizations should participate in industry consortia, standardization efforts, and public consultations. This is particularly important in shaping harmonised standards that will define "state of the art" AI governance. Participation allows organizations to align early, and to help shape governance expectations that are realistic and effective.

External oversight also plays a disciplinary role. It discourages symbolic governance and encourages documented, testable practices. Organizations that treat supervisory engagement as an extension of their own governance build resilience and credibility.

Bringing it all together: The role of the AI Compliance Officer

As AI governance matures into a formal discipline, one role increasingly stands at the center of operational integrity: the AI compliance officer (CO). More than a policy interpreter or documentation reviewer, the CO serves as the organization's internal expert, advisor, and coordinator for ensuring that AI systems are governed across their

full lifecycle. In this final section, we consolidate the themes of this chapter by outlining how the CO enables governance to function – daily, institutionally, and strategically.

Positioning and institutional role

The AI compliance officer (CO) is not a theoretical construct. Rather, they are the organizational anchor for everything described in this chapter. In every aspect of the governance workflow they have a central role, they coordinate with other risk and compliance functions and they are able to perform internal oversight and have the organisation ready for audit. That is the role of the CO: to ensure that governance is not just possible, but practiced.

Positioning the CO properly within the organization is critical. As discussed earlier, governance is inherently cross-functional, spanning legal, risk, data science, IT, and product. The CO must sit at the junction of these domains, not subordinate to any one of them. Like the Data Protection Officer (DPO) under GDPR, the CO needs operational independence: they must have access to leadership and to the systems they oversee, but not be responsible for the development or commercial success of those systems. This ensures they can challenge decisions, escalate risks, and flag governance breakdowns without conflict of interest.

In practice, this means the CO must be embedded in governance structures. For example, they may chair or support internal AI review committees, oversee RACI alignment for AI projects or serve as the coordination point between use case owners, legal counsel, and oversight bodies. They must also be visible across the lifecycle: engaged early at intake, consistently involved in documentation and impact assessment processes, and central to post-deployment monitoring and incident response workflows.

The maturity of this role evolves with the organization. In early stages of the AICMM model, the CO may focus on foundational tasks, building system inventories, setting up intake templates, and creating policy awareness. As governance becomes more integrated, the CO transitions into a strategic role, driving continuous improvement, leading retrospectives, reporting to executive leadership, and preparing for external supervisory engagement.

The CO cannot succeed in isolation. Their effectiveness depends on organizational recognition that governance is not a legal burden, but a form of strategic infrastructure. A well-positioned CO is empowered to say “pause” when systems outpace safeguards, to convene when governance breaks down, and to translate abstract legal or ethical expectations into actionable next steps.

Core functions and responsibilities

The AI compliance officer (CO) is not simply a monitor of rules. They are the operational driver of governance. Across this chapter, we have identified critical governance activities: reviewing AI use cases, coordinating oversight mechanisms, managing post-market monitoring, supporting audit readiness and guiding maturity progression. At its core, the CO function can be understood through six interlocking responsibilities, each corresponding to a specific layer of governance:

1. **Informing and Advising**. The CO is responsible for ensuring the organization understands its obligations under the AI Act, data protection laws, and relevant sectoral rules. But beyond legal literacy, they also translate emerging ethical norms and risk signals into actionable guidance. This includes preparing policy briefings, interpreting regulatory developments, and educating teams during intake and design. Their goal is not only to enforce governance, but to raise internal competence.

2. **Embedding Governance into Workflows**. The CO ensures that AI use case intake, risk classification, FRIA/DPIA processes, and escalation routes are not only defined, but actually used. This often involves designing intake forms, defining documentation minimums, and ensuring that lifecycle checkpoints are respected. The CO's role here is infrastructural: they make governance workable and repeatable.

3. **Monitoring and Escalation**. From continuous oversight to structured incident response, the CO must ensure that systems are not just deployed, but actively stewarded. This includes reviewing monitoring dashboards, verifying that drift thresholds are meaningful, and ensuring that escalation protocols are in place and tested. Where oversight bodies exist, the CO ensures that they are fed the right information at the right time.

4. **Policy Development and Alignment**. The CO plays a coordinating role in developing and updating AI policies—aligning them with internal standards (risk, security, ESG) and external norms (AI Act, ISO standards, ALTAI). This means integrating with existing control frameworks rather than creating redundant ones. The CO ensures that policies are not abstract, but translated into clear procedures and responsibilities.

5. **Audit Preparation and External Engagement**. The CO is the natural lead for preparing audit trails, responding to regulator requests, and maintaining documentation hygiene. They are often the first point of contact for supervisory bodies, external auditors, or certification bodies. But their role is proactive as well: ensuring documentation is structured, impact assessments are current, and use case decisions are traceable.

❻ **Internal Culture and Capability Building**. Governance doesn't function if it's feared or misunderstood. The CO must build a compliance-aware culture, running workshops, internal training, and retrospectives that make governance tangible and collaborative. They are responsible not just for rule-following, but for making governance intelligible and actionable across teams.

Capabilities and learning goals

The effectiveness of an AI compliance officer (CO) rests not only on where they are positioned or what tasks they are assigned, but on what they are capable of doing in practice. Governance is a human function, depending on judgment, communication, pattern recognition, and influence. These are foundational to whether governance succeeds.

This section sets out a capability framework for the CO, grounded in practice but oriented toward professionalization. These five domains reflect not only what the role entails, but what a person must be able to do to carry it out responsibly. They draw from legal theory, ethics, organizational behavior, and risk governance—but are articulated in terms of operational capability, not abstract qualification.

❶ **Deep Understanding of the Ethical and Legal Dimensions of AI Governance.** To govern AI systems responsibly, the CO must be fluent in both binding rules and normative expectations. This means understanding not only the letter of the AI Act, but also the broader ethical questions it seeks to address—bias, explainability, human dignity, and societal impact. The CO must be able to identify when systems raise legal or ethical concerns, even when they technically meet performance requirements. They act as an internal conscience as well as a compliance guide.

❷ **Insight into the Global AI Regulatory Landscape.** The CO must be able to track and interpret developments in other jurisdictions, including influential risk management frameworks, classification models, or sectoral rules in health, finance, or education. This comparative awareness enables organizations to anticipate requirements, benchmark against best practices, and avoid fragmented or duplicative governance approaches.

❸ **Analytical and Problem-Solving Skills for Complex AI Questions.** AI systems generate complex governance problems: composite risks, unclear accountability chains, or ethical dilemmas with no clear precedent. The CO must be able to identify these situations, gather relevant input, and lead structured problem-solving processes. Whether coordinating a cross-functional review or drafting a policy exception, they need to think systemically and act decisively.

4. **Strategic Autonomy and Policy Execution Capability.** Governance cannot depend on case-by-case approvals from senior leadership. The CO must be trusted to act autonomously, leading initiatives, shaping policy, and coordinating execution across teams. This includes developing governance materials, overseeing AI use case intake, defining lifecycle checkpoints, and improving internal processes over time. They must balance strategic vision with operational pragmatism.

5. **Organizational Sensitivity and Constructive Influence.** The CO must operate in real organizations, with competing priorities, legacy systems, and varying levels of awareness. They need to influence without coercion, engaging stakeholders constructively, adapting language to different audiences, and managing resistance with insight and diplomacy. This cultural competence is often what determines whether governance becomes embedded or sidelined.

These capabilities do not emerge from policy documents. They are developed through exposure, reflection, and structured training. Organizations that take AI governance seriously must support COs in acquiring them – not only by formal certification, but by giving the role space to grow into its full strategic potential.

The CO as governance catalyst

In organizations serious about aligning AI with public interest, legal standards, and internal integrity, the CO is what turns principle into procedure, and procedure into practice. They make governance possible not through authority alone, but through structure, continuity, and example. As AI technologies change rapidly and their risks shift, an institutional actor is needed who can convene, interpret, escalate, and adapt. The CO's role is to keep governance responsive.

The CO is also a translator between domains. They speak the language of legal constraints, ethical principles, technical parameters, and organizational priorities. Their role is to surface frictions early—between accuracy and fairness, innovation and accountability, automation and human agency—and to mediate those tensions constructively. This requires not just skill, but trust: the CO must be seen as credible by engineers, lawyers, executives, and auditors alike.

Most of all, the CO sustains the conditions for trust. In environments shaped by uncertainty and complexity, trust is not granted—it is earned, and maintained through transparency, accountability, and repeatable processes. The CO's contribution lies in making those processes visible, explainable, and auditable—not once, but over time.

In this sense, the CO is not simply enforcing governance. They are governing governance: making it possible, making it operational, and ensuring that it matures alongside the systems it is meant to oversee. As AI becomes more systemic, embedded, and consequential, the CO ensures that organizations don't just deploy it efficiently—but govern it accountably, ethically, and with foresight.

Key takeaways

AI governance has moved from a conceptual ambition to an operational imperative. Across this book, we've explored the legal frameworks, ethical principles, and procedural tools that shape responsible AI development and deployment. From the AI Act to lifecycle risk assessments, from transparency obligations to the evolving role of the AI compliance officer, the message is clear: trustworthy AI doesn't emerge by default—it must be structured, sustained, and stewarded. You now hold the language, the frameworks, and the critical mindset needed to do that work well.

As you close this book, recognize your role not just as a reader, but as a practitioner. You are entering a landscape that needs clear thinking, grounded values, and practical leadership. Whether you're advising an executive team, reviewing a high-risk system, designing oversight workflows, or building governance from the ground up – your work matters. AI will continue to evolve, and so will its risks. But so too will the tools, communities, and standards that support its responsible use. Stay connected. Stay curious. And above all, stay committed to shaping AI not just as a technology, but as a force for accountable, ethical, and human-centered progress.

References

CHAPTER 1

1 **Noguchi, T., Hashizume, Y., Moriyama, H., Gauthier, L., Ishikawa, Y., Matsuno, T., & Suganuma, A.** (2018, May). A practical use of expert system" AI-Q" focused on creating training data. In 2018 5th International Conference on Business and Industrial Research (ICBIR) (pp. 73-76). IEEE. 2 **Gailhofer, P., Herold, A., Schemmel, J. P., Scherf, C. S., de Stebelski, C. U., Köhler, A. R., & Braungardt, S.** (2021). The role of artificial intelligence in the European Green Deal. Luxembourg, Belgium: European Parliament. 3 **D.M. Nabirahni, B.R. Evans & A. Persaud,** 'Al-Khwarizmi (algorithm) and the development of algebra', Mathematics Teaching Research Journal 2019, 11, p. 13-17. 4 **Bordot, F.** (2022). Artificial Intelligence, Robots and Unemployment: Evidence from OECD Countries. Journal of Innovation Economics & Management, 37, 117-138. https://doi.org/10.3917/jie.037.0117 5 **Wagner, B. et al.** 'Algorithms and human rights. Study on the human rights dimensions of automated data processing techniques and possible regulatory implications, DGI(2017)12, prepared by the Committee of Experts on internet intermediaries (MSI-NET) for the Council of Europe' (2018) https://rm.coe.int/algorithms-and-human-rights-en-rev/16807956b5 6 **Council of Europe,** 'Algorithms and Human Rights: Study on the human rights dimensions of automated data processing techniques and possible regulatory implications', DGI(2017)12. https://rm.coe.int/algorithms-and-human-rights-en-rev/16807956b5 7 **Dubber, M. D., Pasquale, F., & Das, S. (Eds.).** (2020). The Oxford handbook of ethics of AI. Oxford Handbooks. 8 **Steinhoff, J.** (2023). AI ethics as subordinated innovation network. AI & SOCIETY, 1-13. 9 **Green, B.** (2021). The contestation of tech ethics: A sociotechnical approach to technology ethics in practice. Journal of Social Computing, 2(3), 209-225. 10 **Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... & Vayena, E.** (2018). AI4People – an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. Minds and machines, 28, 689-707. 11 **M. Haenlein & A.Kaplan,** 'A brief history of artificial intelligence: On the past, present, and future of artificial intelligence', California management review 2019, 61.4, p. 5-14. 12 **Floridi, L.** What the Near Future of Artificial Intelligence Could Be. Philos. Technol. 32, 1–15 (2019). https://doi.org/10.1007/s13347-019-00345-y 13 **Turpin, M., Michael, J., Perez, E., & Bowman, S.** (2023). Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. Advances in Neural Information Processing Systems, 36, 74952-74965. 14 **Walsh, T.** (2017). The singularity may never be near. ai Magazine, 38(3), 58-62. 15 **Wang, F. Y., Zhang, J. J., Zheng, X., Wang, X., Yuan, Y., Dai, X., ... & Yang, L**. (2016). Where does AlphaGo go: From church-turing thesis to AlphaGo thesis and beyond. IEEE/CAA Journal of Automatica Sinica, 3(2), 113-120. 16 **Birnbaum, J.** (1997). Pervasive information systems, Communications of the ACM, 40-2, p. 40-41. 17 **Turing, A.M.** (1950). Computing Machinery and Intelligence, 59 MIND 433, 442. 18 **R.S. Boyer e.a., 'In memoriam: Edsger W. Dijkstra 1930 – 2002',** Communications of the ACM 2002, 45, 10, p. 21-22.

19 **Hachey, K. K., Libel, T., & Partington, Z.** (2020). The impact of artificial intelligence on the military profession. Rethinking Military Professionalism for the Changing Armed Forces, 201-211. 20 **Beard, J. M.** (2013). Autonomous weapons and human responsibilities. Geo. J. Int'l L., 45, 617. 21 **Liu, X., Faes, L., Kale, A. U., Wagner, S. K., Fu, D. J., Bruynseels, A., ... & Denniston, A. K.** (2019). A comparison of deep learning performance against healthcare professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. The Lancet Digital Health, 1(6), e271–e297. 22 **Lekadir, K., Frangi, A. F., Porras, A. R., Glocker, B., Cintas, C., Langlotz, C. P., ... & Starmans, M. P.** (2025). FUTURE-AI: International consensus guideline for trustworthy and deployable artificial intelligence in healthcare. bmj, 388. 23 **Campbell, J. F., & Green, K. M.** (2021). Robots as Caretakers. Intersectional Automations: Robotics, AI, Algorithms, and Equity, 169. 24 **Lancaster, K.** (2019). The Robotic Touch: Why there is no good reason to prefer human nurses to carebots. Philosophy in the Contemporary World, 25(2), 88-109. 25 **Shen, C.** (2024). Fair Use, Licensing, and Authors' Rights in the Age of Generative AI. Nw. J. Tech. & Intell. Prop., 22, 157. 26 **Engelfriet, A. & Visser, D.** Beschermt de mijnwerk opt-out mijn werk? Auteursrecht 2024/1. 27 **Pundir, D. S., Jindal, B., Ranga, P., Saini, V., Mahajan, N., & Pandey, K. S.** (2025). The Rise of Artificial Intelligence in Intellectual Property Law: Patentability and Copyright Issues. Metallurgical and Materials Engineering, 31(4), 40-43. 28 **Lazar, S.** (2024). Automatic authorities: AI, legitimacy, and democratic erosion. Preprint available at arXiv:2404.05990. 29 **Abashidze, A. K., Ilyashevich, M., & Latypova, A.** (2022). Artificial intelligence and space law. J. Legal Ethical & Regul. Issues, 25, 1. 30 **Graham, T., Thangavel, K., & Martin, A. S.** (2024). Navigating AI-lien Terrain: Legal liability for artificial intelligence in outer space. Acta Astronautica, 217, 197-207.

CHAPTER 2

1 **Pošćić, A., & Martinović, A. (2021).** Towards a Regulatory Framework for Artificial Intelligence-an EU Approach. In *Contemporary Economic and Business Issues* (pp. 49-62). Sveučilište u Rijeci, Ekonomski fakultet. 2 **Sharma, R., Lopes de Sousa Jabbour, A. B., Jain, V., & Shishodia, A.** (2022). The role of digital technologies to unleash a green recovery: Pathways and pitfalls to achieve the European Green Deal. *Journal of Enterprise Information Management*, 35(1), 266-294. 3 **Ulnicane, I., Knight, W., Leach, T., Stahl, B. C., & Wanjiku, W. G.** (2021). Framing governance for a contested emerging technology: Insights from AI policy. *Policy and Society*, 40(2), 158–177. https://doi. org/10.1080/14494035.2020.1855800 4 **Floridi, L.** (2021). Establishing the rules for building trustworthy AI. *Ethics, Governance, and Policies in Artificial Intelligence*, 41-45. 5 **Justo-Hanani, R.** (2022). The politics of Artificial Intelligence regulation and governance reform in the European Union. *Policy Sciences*, 55(1), 137-159. 6 **Bradford, A.** (2020). *The Brussels effect: How the European Union rules the world*. *Oxford University Press*, USA. 7 **Tartaro, A.** (2023). Regulating by standards: current progress and main challenges in the standardisation of Artificial Intelligence in support of the AI Act. *European Journal of Privacy Law & Technologies*, (1). 8 **De Vries, S. A.** (2013). Balancing fundamental rights with economic freedoms according to the European Court of Justice. *Utrecht L. Rev.*, 9, 169. 9 **Zhou, C., Li, Q., Li, C., Yu, J., Liu, Y., Wang, G., ... & Sun, L.** (2024). A comprehensive survey on pretrained foundation models: A history from bert to chatgpt. I*nternational Journal of Machine Learning and Cybernetics*, 1-65. 10 **HERNANDEZ, D. and BROWN, T.** Measuring the algorithmic efficiency of neural networks. *arXiv preprint arXiv*:2005.04305. 2020. 11 **Ranchordas, S.** (2021). Experimental regulations for AI: sandboxes for morals and mores. *University of Groningen Faculty of Law Research Paper*, (7). 12 **Radclyffe, C., Ribeiro, M., & Wortham, R. H.** (2023). The assessment list for trustworthy artificial intelligence: A review and recommendations. *Frontiers in Artificial Intelligence*, 6, 1020592.

CHAPTER 3

1 **Haag, S., & Eckhardt, A.** (2017). Shadow it. *Business & Information Systems Engineering*, 59, 469-473. 2 **Scantamburlo, T., Falcarin, P., Veneri, A., Fabris, A., Gallese, C., Billa, V., ... & Marcuzzi, F.** (2024, April). Software Systems Compliance with the AI Act: Lessons Learned from an International Challenge. In *Proceedings of the 2nd International Workshop on Responsible AI Engineering* (pp. 44-51).

3 **Hupont, I., Fernández-Llorca, D., Baldassarri, S.** et al. Use case cards: a use case reporting framework inspired by the European AI Act. *Ethics Inf Technol* 26, 19 (2024). https://doi.org/10.1007/s10676-024-09757-7. 4 **Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T.** (2019, January). Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 220-229) 5 **Floridi, L., Holweg, M., Taddeo, M., Amaya Silva, J., Mökander, J., & Wen, Y.** (2022). CapAI-A procedure for conducting conformity assessment of AI systems in line with the EU artificial intelligence act. *Available at SSRN 4064091*. 6 **Simonetta, A., & Paoletti, M. C.** (2024). ISO/IEC Standards and Design of an Artificial Intelligence System. IWESQ 2024: *6th International Workshop on Experience with SQuaRE family and its Future Direction*, 1-6. 7 **Mantelero, A.**: AI and big data: a blueprint for a human rights, social and ethical impact assessment. Comput. Law Secur. Rev. 34(4), 754–772 (2018). 8 **Mantelero, A.** (2024). The Fundamental Rights Impact Assessment (FRIA) in the AI Act: Roots, legal obligations and key elements for a model template. *Computer Law & Security Review*, 54, 106020. 9 **Harer, J.** (2023). Post-Market Surveillance and Vigilance on the European Market. In *Medical Devices and In Vitro Diagnostics: Requirements in Europe* (pp. 1-39). Cham: Springer International Publishing. 10 **Mökander, J., Axente, M., Casolari, F., & Floridi, L.** (2022). Conformity assessments and post-market monitoring: a guide to the role of auditing in the proposed European AI regulation. *Minds and Machines*, 32(2), 241-268. 11 **Chittala, S.** (2024). AIOps in Action: Automating AI Deployment and Management of Large Language Models for Scalable and Ethical Operations. *IJFMR*, 6. 12 **Almada, M., & Petit, N.** (2025). The EU AI Act: Between the rock of product safety and the hard place of fundamental rights. *Common market law review*, 62(1). 13 **Hajnal, Z.** (2020). Current challenges of European market surveillance regarding products sold online. *Public Goods Govern*, 5, 1-8. 14 **Stahl, B. C., Rodrigues, R., Santiago, N., & Macnish, K.** (2022). A European Agency for Artificial Intelligence: Protecting fundamental rights and ethical values. *Computer Law & Security Review*, 45, 105661. 15 **Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S.** (2021). Conceptualizing AI literacy: An exploratory review. *Computers and Education: Artificial Intelligence*, 2, 100041. 16 **Long, D., & Magerko, B.** (2020, April). What is AI literacy? Competencies and design considerations. In *Proceedings of the 2020 CHI conference on human factors in computing systems* (pp. 1-16).

CHAPTER 4

1 **Takayama, L.** (2015). Telepresence and apparent agency in human–robot interaction. In S. S. Sundar (Ed.), *The handbook of the psychology of communication technology* (pp. 160–175). Malden, MA: Wiley Blackwell. 2 **J. Greenwood, A. Seshadri & M. Yorukoglu**, 'Engines of liberation', *The Review of Economic Studies* 2005, 72/1, p. 109-133. 3 **Sarter, N. B., Woods, D. D., and Billings, C. E.** (1997). "Automation surprises," in *Handbook of human factors and ergonomics*, 2nd Edn, ed. G. Salvendy (New York, NY: Wiley), 1926–1943. 4 **Dekker, S. W. A., and Woods, D. D.** (2002). MABA-MABA or abracadabra? Progress on human-automation co-ordination. *Cogn. Technol. Work* 4, 240–244. 5 **Endsley, M. R.** (1999). Level of automation effects on performance, situation awareness and workload in a dynamic control task. Ergonomics 42, 462–492. 6 **Pagliari, M., Chambon, V., & Berberian, B.** (2022). What is new with Artificial Intelligence? Human–agent interactions through the lens of social agency. Frontiers in Psychology, 13, 954444. 7 **Barlas, Z., Hockley, W. E., & Obhi, S. S.** (2017). The effects of freedom of choice in action selection on perceived mental effort and the sense of agency. *Acta psychologica*, 180, 122-129. 8 **Cheung, A. S., & Chen, Y.** (2022). From datafication to data state: Making sense of China's social credit system and its implications. *Law & Social Inquiry*, 47(4), 1137-1171. 9 **Vokey, J. R., & Read, J. D.** (1985). Subliminal messages: Between the devil and the media. *American psychologist*, 40(11), 1231. 10 **Neuwirth, R. J.** (2022). *The EU artificial intelligence act: regulating subliminal AI systems*. Taylor & Francis. 11 **Silver, C. A., Tatler, B. W., Chakravarthi, R., and Timmermans, B.** (2020). Social agency as a continuum. *Psycho. Bull.* Rev. 28, 434–453. doi: 10.3758/s13423-020-01845-1 12 **Vasconcelos, H., Jörke, M., Grunde-McLaughlin, M., Gerstenberg, T., Bernstein, M. S., & Krishna, R.** (2023). Explanations can reduce overreliance on ai systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1), 1-38. 13 **Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., ... & Weld, D.** (2021, May). Does the whole exceed its parts? the effect of AI explanations on complementary team performance. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (pp. 1-16). 14 **Vasconcelos, H., Jörke, M., Grunde-McLaughlin, M., Gerstenberg, T., Bernstein, M. S., & Krishna, R.** (2023). Explanations can reduce over-reliance on ai systems during decision-making. Proceedings of the ACM on Human-Computer Interaction, 7(CSCW1), 1-38. 15 **Buçinca, Z., Malaya, M. B., & Gajos, K. Z.** (2021). To trust or to think: cognitive forcing functions can reduce over-reliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1), 1-21. 16 **Donahoe, E., & Metzger, M. M.** (2019). Artificial intelligence and human rights. *J. Democracy*, 30, 115. 17 **Abedin, B., Meske, C., Junglas, I., Rabhi, F., & Motahari-Nezhad, H. R.** (2022). Designing and managing human-AI interactions. *Information Systems Frontiers*, 24(3), 691-697. 18 **Schuetz, S., & Venkatesh, V.** (2020). The rise of human machines: How cognitive computing systems challenge

assumptions of user-system interaction. Journal of the Association for Information Systems, 21(2), 460-482. 19 **Kędzierski, J., Kaczmarek, P., Dziergwa, M., & Tchoń, K.** (2015). Design for a robotic companion. International journal of humanoid robotics, 12(01), 1550007. 20 **Dillon, S.** (2020). The Eliza effect and its dangers: From demystification to gender critique. Journal for Cultural Research, 24(1), 1-15. 21 **Belk, R.** (2022). Artificial emotions and love and sex doll service workers. Journal of Service Research, 25(4), 521-536. 22 **Richardson, K.** (2016). Sex robot matters: slavery, the prostituted, and the rights of machines. *IEEE Technology and Society Magazine*, 35(2), 46-53. 23 **Xie, T., & Pentina, I.** (2022). Attachment theory as a framework to understand relationships with social chatbots: a case study of Replika. 24 **Cofer, D.** (2021, October). Unintended behavior in learning-enabled systems: detecting the unknown unknowns. In 2021 *IEEE/AIAA 40th Digital Avionics Systems Conference (DASC)* (pp. 1-7). IEEE. 25 **Arnold, T., & Scheutz, M.** (2018). The "big red button" is too late: an alternative model for the ethical evaluation of AI systems. *Ethics and Information Technology*, 20, 59-69.

CHAPTER 5

1 **Papakonstantinou, V.** (2022). Cybersecurity as praxis and as a state: The EU law path towards acknowledgement of a new right to cybersecurity?. *Computer Law & Security Review, 44*, 105653. 2 **Chiara, P. G.** (2022). The IoT and the new EU cybersecurity regulatory landscape. *International Review of Law, Computers & Technology, 36*(2), 118-137. 3 **Clim, A., Toma, A., Zota, R. D., & Constantinescu, R.** (2022). The Need for Cybersecurity in Industrial Revolution and Smart Cities. *Sensors, 23*(1), 120. 4 **Bae, H., Jang, J., Jung, D., Jang, H., Ha, H., Lee, H., & Yoon, S.** (2018). Security and privacy issues in deep learning. *arXiv preprint arXiv:1807.11655*. 5 **Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., ... & Song, D.** (2018). Robust physical-world attacks on deep learning visual classification. *In Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1625-1634). 6 **Ozlati, S., & Yampolskiy, R.** (2017, March). The formalization of AI risk management and safety standards. *In Workshops at the Thirty-First AAAI Conference on Artificial Intelligence*. 7 **Taherdoost, H.** (2022). Understanding cybersecurity frameworks and information security standards – a review and comprehensive overview. *Electronics, 11*(14), 2181. 8 **De Haes, S., Van Grembergen, W., & Debreceny, R. S.** (2013). COBIT 5 and enterprise governance of information technology: Building blocks and research opportunities. *Journal of Information Systems*, 27(1), 307-324. 9 **Dabade, T. D.** (2012). Information technology infrastructure library (ITIL). *In Proceedings of the 4th National Conference* (pp. 25-26). 10 **Dunn Cavelty, M., & Smeets, M.** (2023). Regulatory cybersecurity governance in the making: The formation of ENISA and its struggle for epistemic authority. *Journal of European Public Policy*, 30(7), 1330-1352. 11 **Siau, K., & Wang, W.** (2018). Building trust in artificial intelligence, machine learning, and robotics. *Cutter business technology journal*, 31(2), 47-53. 12 **Hubbard, D. W.** (2009). *The Failure of Risk Management*: Why It's Broken and How to Fix It. John Wiley & Sons. 13 **Arora, A. S., Changotra, R., & Rajput, H.** (2021). to Quantitative Risk Assessment Methodologies. *Bow Ties in Process Safety and Environmental Management*: *Current Trends and Future Perspectives,* 211. 14 **Lior, A.** (2022). Insuring AI: The role of insurance in artificial intelligence regulation. *Harvard Journal of Law and Technology*, 1. 15 **Leffingwell, D., Meissner, M., & Langenfeld, C.** (2011). Agile software development with verification and validation in high assurance and regulated environments. *Rally Software Development Corp.* 16 **Wang, L., Zhang, X., Su, H., & Zhu, J.** (2023). A comprehensive survey of continual learning: Theory, method and application. *arXiv preprint arXiv:2302.00487.* 17 **Aljundi, R., Kelchtermans, K., & Tuytelaars, T.** (2019). Task-free continual learning. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 11254-11263).

CHAPTER 6

1 **Westin, A. F.** (1967). *Privacy and freedom*. Atheneum. 2 **Riccardi, J. L.** (1983). The German federal data protection act of 1977: Protecting the right to privacy. Boston College International and Comparative Law Review, 6, 243. 3 **Bennett, C. J.** (2018). The European General Data Protection Regulation: An instrument for the globalization of privacy standards? Information Polity, 23(2), 239-246. 4 **Schwartz, P. M., & Solove, D. J.** (2011). The PII problem: Privacy and a new concept of personally identifiable information. *NYU Law Review*, 86, 1814. 5 **Barrett, C.** (2019). Are the EU GDPR and the California CCPA becoming the de facto global standards for data privacy and protection?. *Scitech Lawyer, 15*(3), 24-29. 6 **Haenlein, M., & Kaplan, A.** (2019). A brief history of artificial intelligence: On the past, present, and future of artificial intelligence. *California management review, 61*(4), 5-14. 7 **Zuboff, S.** (2015). Big other: surveillance capitalism and the prospects of an information civilization. *Journal of information technology, 30*(1), 75-89. 8 **Kalluri, P. R., Agnew, W., Cheng, M., Owens, K., Soldaini, L., & Birhane, A.** (2023). The Surveillance AI Pipeline. *arXiv preprint arXiv:2309.15084*. 9 **CNIL** (2023). AI how-to sheets. Retrieved from: https://www.cnil.fr/en/ai-how-sheets 10 **Cavoukian, A.** (2009). Privacy by design: The 7 foundational principles. *Information and privacy commissioner of Ontario, Canada, 5*, 12. 11 **Labadie, C., & Legner, C.** (2020). Personal Data Protection Inside and Out: Integrating Data Protection Requirements in the Data Lifecycle. *Enterprise Modelling and Information Systems Architectures (EMISAJ)*, 15, 9-1. 12 **Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K.** (2021). Datasheets for datasets. *Communications of the ACM*, 64(12), 86-92. 13 **Pushkarna, M., Zaldivar, A., & Kjartansson, O.** (2022, June). Data cards: Purposeful and transparent dataset documentation for responsible ai. *In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1776-1826). 14 **Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T.** (2019, January). Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 220-229). 15 **Chinta, S. V., Wang, Z., Yin, Z., Hoang, N., Gonzalez, M., Quy, T. L., & Zhang, W.** (2024). FairAIED: Navigating fairness, bias, and ethics in educational AI applications. *arXiv preprint arXiv:* 2407.18745. 16 **Zaharia, M., Chen, A., Davidson, A., Ghodsi, A., Hong, S. A., Konwinski, A., ... & Zumar, C.** (2018). Accelerating the machine learning lifecycle with MLflow. *IEEE Data Eng. Bull.*, 41(4), 39-45. 17 **Tarnas, C., & Nguyen, V.** (2023). *MLOps Engineering at Scale*. O'Reilly Media. 18 **Zhang, D., Finckenberg-Broman, P., Hoang, T., Pan, S., Xing, Z., Staples, M., & Xu, X.** (2024). Right to be forgotten in the era of large language models: Implications, challenges, and solutions. *AI and Ethics*, 1-10. 19 **Rosati, E.** (2019). Copyright as an obstacle or an enabler? A European perspective on text and data mining and its role in the development of AI creativity. *Asia Pacific Law Review*, 27(2), 198-217. 20 **Engelfriet, A. and Visser, D**. Werkt de

mijnwerk opt-out voor mijn werk? *Auteursrecht*. 2024. Vol. 1. 21 **Omelina, L., Goga, J., Pavlovicova, J., Oravec, M., & Jansen, B.** (2021). A survey of iris datasets. *Image and Vision Computing*, 108, 104109. 22 **Ahler, D. J., Roush, C. E., & Sood, G.** (2019). The micro-task market for lemons: Data quality on Amazon's Mechanical Turk. *Political Science Research and Methods*, 1-20. 23 N**guyen, T., Ilharco, G., Wortsman, M., Oh, S., & Schmidt, L.** (2022). Quality not quantity: On the interaction between dataset design and robustness of clip. *Advances in Neural Information Processing Systems*, 35, 21455-21469. See also Birhane, A., Prabhu, V., Han, S., & Boddeti, V. N. (2023). On Hate Scaling Laws For Data-Swamps. arXiv preprint arXiv:2306.13141. 24 **Yang, K., Qinami, K., Fei-Fei, L., Deng, J., & Russakovsky, O.** (2020, January). Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. *In Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 547-558). 25 **M. Ebers e.a.,** 'The European Commission's Proposal for an Artificial Intelligence Act - A Critical Assessment by Members of the Robotics and AI Law Society (RAILS)', J 2021, 4, p. 589- 603 26 **Budach, L., Feuerpfeil, M., Ihde, N., Nathansen, A., Noack, N., Patzlaff, H., ... & Harmouch, H.** (2022). The effects of data quality on machine learning performance. *arXiv preprint* arXiv:2207.14529. 27 **Liang, W., Tadesse, G. A., Ho, D., Fei-Fei, L., Zaharia, M., Zhang, C., & Zou, J.** (2022). Advances, challenges and opportunities in creating data for trustworthy AI. Nature Machine Intelligence, 4(8), 669-677. 28 **Munappy, A., Bosch, J., Olsson, H. H., Arpteg, A., & Brinne, B.** (2019, August). Data management challenges for deep learning. In 2019 45th Euromicro Conference on Software Engineering and Advanced Applications (SEAA) (pp. 140-147). IEEE.

CHAPTER 7

1 **Larsson, S., & Heintz, F.** (2020). Transparency in artificial intelligence. *Internet Policy Review, 9*(2). 2 **Jobin, A., Ienca, M., & Vayena, E.** (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*(9), 389–399. 3 **Pasquale, F.** (2015). *The black box society: The secret algorithms that control money and information.* Harvard University Press. 4 **Dubber, M. D., Pasquale, F., & Das, S.** (Eds.). (2020). *The Oxford handbook of ethics of AI.* Oxford Handbooks. 5 **Lu, J., Liu, A., Dong, F., Gu, F., Gama, J., & Zhang, G.** (2018). Learning under concept drift: A review. *IEEE transactions on knowledge and data engineering, 31*(12), 2346-2363. 6 **Kreuzberger, D., Kühl, N., & Hirschl, S.** (2023). Machine learning operations (mlops): Overview, definition, and architecture. *IEEE Access.* 7 **Mora-Cantallops, M., Sánchez-Alonso, S., García-Barriocanal, E., & Sicilia, M. A.** (2021). Traceability for trustworthy ai: A review of models and tools. *Big Data and Cognitive Computing, 5*(2), 20. 8 **Plesser, H. E.** (2018). Reproducibility vs. replicability: a brief history of a confused terminology. *Frontiers in neuroinformatics, 11*, 76. 9 Kaplan, A. (2020). Artificial intelligence, social media, and fake news: Is this the end of democracy. *IN MEDIA & SOCIETY, 149.* 10 **Lancaster, T.** (2023). Artificial intelligence, text generation tools and ChatGPT–does digital watermarking offer a solution?. *International Journal for Educational Integrity, 19*(1), 10. 11 **Zhu, J., He, P., Fu, Q., Zhang, H., Lyu, M. R., & Zhang, D.** (2015, May). Learning to log: Helping developers make informed logging decisions. In *2015 IEEE/ACM 37th IEEE International Conference on Software Engineering* (Vol. 1, pp. 415-425). IEEE. 12 **Bosch, N., & Bosch, J.** (2020). Software logging for machine learning. *arXiv preprint arXiv:2001.10794.* 13 **Confalonieri, R., Coba, L., Wagner, B., & Besold, T. R.** (2021). A historical perspective of explainable Artificial Intelligence. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 11*(1), e1391. 14 **Xu, F., Uszkoreit, H., Du, Y., Fan, W., Zhao, D., & Zhu, J.** (2019). Explainable AI: A brief survey on history, research areas, approaches and challenges. In *Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II 8* (pp. 563-574). Springer International Publishing. 15 **Bücker, M., Szepannek, G., Gosiewska, A., & Biecek, P.** (2022). Transparency, auditability, and explainability of machine learning models in credit scoring. *Journal of the Operational Research Society, 73*(1), 70-90. 16 **McCoy, L. G., Brenna, C. T., Chen, S. S., Vold, K., & Das, S.** (2022). Believing in black boxes: machine learning for healthcare does not need explainability to be evidence-based. *Journal of clinical epidemiology, 142*, 252-257. 17 **Miller, T.** (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence, 267*, 1-38. 18 **Edwards, L., & Veale, M.** (2018). Enslaving the algorithm: From a "right to an explanation" to a "right to better decisions"?. *IEEE Security & Privacy*, 16(3), 46-54.. 19 **Zonneveldt, S., Korb, K., & Nicholson, A.** (2010). Bayesian network classifiers for the German credit data. *Bayesian-intelligence. com/*

publications. 20 **Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F.** (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion, 58*, 82-115. 21 **Ribeiro, M. T., Singh, S., & Guestrin, C.** (2016, August). "Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135-1144). 22 **Sundararajan, M., Taly, A., & Yan, Q. (**2017, July). Axiomatic attribution for deep networks. In International conference on machine learning (pp. 3319-3328). PMLR.. 23 **Wachter, S., Mittelstadt, B., & Russell, C.** (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech., 31*, 841. 24 **Lipton, Z. C.** (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3), 31-57.. 25 **Waldman, A. E.** (2019). Power, process, and automated decision-making. *Fordham L. Rev., 88,* 613.. 26 **Binns, R., & Veale, M.** (2021). Is that your final decision? Multi-stage profiling, selective effects, and Article 22 of the GDPR. *International Data Privacy Law, 11*(4), 319-332. 27 **Aloisi, A., & Potocka-Sionek, N.** (2022). De-gigging the labour market? An analysis of the 'algorithmic management'provisions in the proposed Platform Work Directive. *An Analysis of the 'Algorithmic Management'Provisions in the Proposed Platform Work Directive (July 21, 2022)*. 28 **Burrell, J.** (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big data & society,* 3(1), 2053951715622512. 29 **De Groot, A.** (2023). Care to explain? *A critical epistemic in/justice-based analysis of legal explanation obligations and ideals for "AI"-infused times* [Ph.D. dissertation], Tilburg Univ., Tilburg, Netherlands.. 30 **Almada, M**. (2019). Human intervention in automated decision-making: Toward the construction of contestable systems. In *Proceedings of the Seventeenth International Conference on artificial intelligence and law* (pp. 2-11).

CHAPTER 8

1 **Carey, A. N., & Wu, X.** (2023). The statistical fairness field guide: perspectives from social and formal sciences. *AI and Ethics*, *3*(1), 1-23. 2 **Hellström, T., Dignum, V., & Bensch, S.** (2020). Bias in Machine Learning – What is it Good for?. *arXiv preprint arXiv:2004.00686*. 3 **Miller, K.** (2020). A matter of perspective: Discrimination, bias, and inequality in ai. In *Legal regulations, implications, and issues surrounding digital data* (pp. 182-202). IGI Global. 4 **Osoba, O. A., Welser IV, W., & Welser, W.** (2017). *An intelligence in our image: The risks of bias and errors in artificial intelligence*. Rand Corporation. 5 **Caliskan, A., Bryson, J. J., & Narayanan, A.** (2017). Semantics derived automatically from language corpora contain human-like biases. *Science, 356*(6334), 183-186. 6 **Papakyriakopoulos, O., & Mboya, A. M.** (2023). Beyond algorithmic bias: a socio-computational interrogation of the google search by image algorithm. *Social Science Computer Review*, *41*(4), 1100-1125. 7 **Datta, A., Tschantz, M. C., & Datta, A.** (2014). Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination. *arXiv preprint arXiv:1408.6491*. 8 **Fleisig, E., Smith, G., Bossi, M., Rustagi, I., Yin, X., & Klein, D.** (2024). Linguistic Bias in ChatGPT: Language models reinforce dialect discrimination. *arXiv preprint arXiv:2406.08818*. 9 **Kuśmierczyk, M.** (2022). Algorithmic Bias in the Light of the GDPR and the Proposed AI Act. *In) equality. Faces of modern Europe", Wydawnictwo Centrum Studiów Niemieckich i Europejskich im. Willy'ego Brandta, Wrocław*. 10 **Chi, N., Lurie, E., & Mulligan, D. K.** (2021, July). Reconfiguring diversity and inclusion for AI ethics. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 447-457). 11 **Wirth, R., & Hipp, J.** (2000, April). CRISP-DM: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining* (Vol. 1, pp. 29-39). 12 **Martínez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernández-Orallo, J., Kull, M., Lachiche, N., ... & Flach, P.** (2019). CRISP-DM twenty years later: From data mining processes to data science trajectories. *IEEE Transactions on Knowledge and Data Engineering, 33*(8), 3048-3061. 13 **Stoyanovich, J., & Howe, B.** (2019). Nutritional labels for data and models. *A Quarterly bulletin of the Computer Society of the IEEE Technical Committee on Data Engineering*, 42(3). 14 **Pagano, T. P., Loureiro, R. B., Lisboa, F. V., Peixoto, R. M., Guimarães, G. A., Cruz, G. O., ... & Nascimento, E. G.** (2023). Bias and unfairness in machine learning models: a systematic review on datasets, tools, fairness metrics, and identification and mitigation methods. *Big data and cognitive computing*, 7(1), 15. 15 **van Bekkum, M.** (2025). Using sensitive data to de-bias AI systems: Article 10 (5) of the EU AI act. Computer Law & Security Review, 56, 106115. 16 **Bittenbinder, S., Müller, C., & Tuncer, Z.** (2023). European Accessibility Act-Practice-based approaches to meeting accessibility requirements. *Mensch und Computer 2023 - Workshopband*. 17 **Martínez-Normand, L., & Pluke, M.** (2014). A decision-tree approach for the applicability of the accessibility standard EN 301 549. In *Computers*

Helping People with Special Needs: 14th International Conference, ICCHP 2014, Paris, France, July 9-11, 2014, Proceedings, Part II 14 (pp. 295-302). Springer International Publishing. 18 **Stephanidis, C., Salvendy, G., Antona, M., Chen, J. Y., Dong, J., Duffy, V. G., ... & Zhou, J.** (2019). Seven HCI grand challenges. *International Journal of Human–Computer Interaction, 35*(14), 1229-1269. 19 **Stephanidis, C.** (2021). Design for all in digital technologies. *Handbook of human factors and ergonomics*, 1187-1215. 20 **Altinier, A., Oncins, E., Sauberer, G., & Mehigan, T.** (2022, August). Demystifying digital accessibility and fostering inclusive mindsets. Compliance with the *European standard for digital accessibility EN 301 549*. *In European Conference on Software Process Improvement* (pp. 595-609). Cham: Springer International Publishing. 21 **Reich, K., & Petter, C.** (2009). eInclusion, eAccessibility and design for all issues in the context of European computer-based assessment. *The transition to computer-based assessment. New approaches to skills assessment and implications for large-scale testing*, 68-73. 22 **Lee, M. K., Kusbit, D., Kahng, A., Kim, J. T., Yuan, X., Chan, A., ... & Procaccia, A. D.** (2019). WeBuildAI: Participatory framework for algorithmic governance. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1-35. 23 **Delgado, F., Yang, S., Madaio, M., & Yang, Q.** (2021). Stakeholder Participation in AI: Beyond" Add Diverse Stakeholders and Stir". *arXiv preprint arXiv:2111.01122*. 24 **Wong, R. Y., Madaio, M. A., & Merrill, N.** (2023). Seeing like a toolkit: How toolkits envision the work of AI ethics. *Proceedings of the ACM on Human-Computer Interaction, 7*(CSCW1), 1-27. 25 **Bell, A., Nov, O., & Stoyanovich, J.** (2023). Think about the stakeholders first! Toward an algorithmic transparency playbook for regulatory compliance. *Data & Policy, 5*, e12.

CHAPTER 9

1 **Bashir, N.; Donti, P.; Cuff, J.; Sroka, S.; Ilic, M.; Sze, V.; Delimitrou, C.; Olivetti, E.** The Climate and Sustainability Implications of Generative AI. *MIT Explor. Gener. AI 2024*, 1–45. 2 **George, A. S., George, A. H., & Martin, A. G.** (2023). The environmental impact of ai: A case study of water consumption by chat gpt. Partners Universal International Innovation Journal, 1(2), 97-104. 3 **Wang, P., Zhang, L. Y., Tzachor, A., & Chen, W. Q.** (2024). E-waste challenges of generative artificial intelligence. *Nature Computational Science*, 1-6. 4 **Meitei, A. J., Rai, P., & Rajkishan, S. S.** (2025). Application of AI/ML techniques in achieving SDGs: a bibliometric study. *Environment, Development and Sustainability*, 27(1), 281-317. 5 **Aloisi, A.** (2024). Regulating algorithmic management at work in the European Union: Data protection, non-discrimination and collective rights. *International Journal of Comparative Labour Law and Industrial Relations*, 40(1). 6 **Pantanowitz, L., Pearce, T., Abukhiran, I., Hanna, M., Wheeler, S., Soong, T. R., ... & Rashidi, H. H.** (2024). Non-generative artificial intelligence (AI) in medicine: advancements and applications in supervised and unsupervised machine learning. *Modern Pathology*, 100680. 7 **Shanmugam, D., Agrawal, M., Movva, R., Chen, I. Y., Ghassemi, M., Jacobs, M., & Pierson, E.** (2024). *Generative AI in medicine*. arXiv preprint arXiv:2412.10337. 8 **Balogun, E., Dcosta, D., Boch, A., & Luetge, C.** (2024). Exploring key stakeholders' perspectives on integrating the EU AI Act with the MDR for certifying AI medical devices. *AI and Ethics*, 1-15. 9 **Rademakers, F. E., Biasin, E., Bruining, N., Caiani, E. G., Davies, R. H., Gilbert, S. H., ... & Fraser, A. G.** (2025). CORE-MD clinical risk score for regulatory evaluation of artificial intelligence-based medical device software. *npj Digital Medicine*, 8(1), 90. 10 **Rosenbacke, R., Melhus, Å., McKee, M., & Stuckler, D.** (2024). How Explainable Artificial Intelligence Can Increase or Decrease Clinicians' Trust in AI Applications in Health Care: Systematic Review. *JMIR AI, 3*, e53207. 11 **Liu, T., & Duan, Y.** (2024). Beware the self-fulfilling prophecy: enhancing clinical decision-making with AI. *Critical Care*, 28(1), 276. 12 **Scott, I. A., Van Der Vegt, A., Lane, P., McPhail, S., & Magrabi, F.** (2024). Achieving large-scale clinician adoption of AI-enabled decision support. *BMJ Health & Care Informatics*, 31(1), e100971. 13 **Natali, C., Marconi, L., Dias Duran, L. D., Miglioretti, M., & Cabitza, F.** (2025). AI-Induced Deskilling in Medicine: A Mixed Method Literature Review for Setting a New Research Agenda. *Available at SSRN 5166364*. 14 **Akingbola, A., Adeleke, O., Idris, A., Adewole, O., & Adegbesan, A.** (2024). Artificial intelligence and the dehumanization of patient care. *Journal of Medicine, Surgery, and Public Health*, 3, 100138. 15 **Sætra, H. S.** (2021). A Framework for Evaluating and Disclosing the ESG Related Impacts of AI with the SDGs. *Sustainability*, 13(15), 8503. 16 **Badawy, W.** (2025). Algorithmic sovereignty and democratic resilience: rethinking AI governance in the age of generative AI. *AI and Ethics*, 1-8.

CHAPTER 10

1 **Yeung, K.** (2020). Recommendation of the council on artificial intelligence (OECD). *International legal materials, 59*(1), 27-34. 2 **Busuioc, M.** (2021). Accountable artificial intelligence: Holding algorithms to account. Public Administration Review, 81(5), 825-836. 3 **Bovens, M.** (2007). Analysing and Assessing Public Accountability: A Conceptual Framework. *European Law Journal*, 13(4), 447–468. 4 **Centre for Information Policy Leadership.** (2024). *Building Accountable AI Programs*. Washington, DC: CIPL. 5 **Organisation for Economic Co-operation and Development.** (2023). *Advancing Accountability in AI: Lessons from Practice. OECD Digital Economy Papers*, No. 336. Paris: OECD Publishing 6 **Helmer, L., Martens, C., Wegener, D., Becker, D., Akila, M., & Abbas, S.** (2024). *Towards Trustworthy AI Engineering: A Case Study on Integrating an AI Audit Catalog into MLOps Processes.* In Proceedings of the 2024 International Workshop on Responsible AI Engineering (RAIE '24), April 16, Lisbon, Portugal. ACM. 7 **Murikah, W., Nthenge, J. K., & Musyoka, F. M.** (2024). Bias and ethics of AI systems applied in auditing-A systematic review. *Scientific African*, e02281. 8 **Wasil, A. R., Clymer, J., Krueger, D., Dardaman, E., Campos, S., & Murphy, E. R.** (2024). Affirmative safety: An approach to risk management for high-risk AI. *arXiv preprint arXiv:2406.15371*. 9 **Morley, J., Floridi, L., Kinsey, L., & Elhalal, A.** (2020). From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices. *Science and engineering ethics*, 26(4), 2141-2168. 10 **Mäntymäki, M., Minkkinen, M., Birkstedt, T., & Viljanen, M.** (2022). Putting AI ethics into practice: The hourglass model of organizational AI governance. *arXiv preprint arXiv:2206.00335*. 11 **Fanni, R., Steinkogler, V. E., Zampedri, G., & Pierson, J.** (2023). Enhancing human agency through redress in Artificial Intelligence Systems. AI & society, 38(2), 537-547. 12 **Ala-Pietilä, P., Bonnet, Y., Bergmann, U., Bielikova, M., Bonefeld-Dahl, C., Bauer, W., ... & Van Wynsberghe, A.** (2020). *The assessment list for trustworthy artificial intelligence (ALTAI)*. European Commission. 13 **Borg, M., Bronson, J., Christensson, L., Olsson, F., Lennartsson, O., Sonnsjö, E., ... & Karsberg, M.** (2021, June). Exploring the assessment list for trustworthy ai in the context of advanced driver-assistance systems. In *2021 IEEE/ACM 2nd International Workshop on Ethics in Software Engineering Research and Practice (SEthics) (pp. 5-12)*. IEEE. 14 **Golpayegani, D., Pandit, H. J., & Lewis, D.** (2022, December). Comparison and Analysis of 3 Key AI Documents: EU's Proposed AI Act, Assessment List for Trustworthy AI (ALTAI), and ISO/IEC 42001 AI Management System. In *Irish Conference on Artificial Intelligence and Cognitive Science* (pp. 189-200). Cham: Springer Nature Switzerland.

CHAPTER II

1 **Norton, L. W.** (2025). Artificial intelligence and organizational strategy: Ethical and governance implications. *Consulting Psychology Journal*. 2 **Mueller, M. L.** (2025). It's just distributed computing: Rethinking AI governance. *Telecommunications Policy*, 102917. 3 **Batool, A., Zowghi, D., & Bano, M.** (2025). AI governance: a systematic literature review. *AI and Ethics*, 1-15. 4 **Manda, V. K., Christy, V., & Jitta, M. R.** (2025). Ethical AI and Decision-Making in Management Leadership. In *Ethical Dimensions of AI Development* (pp. 197-226). IGI Global. 5 **Project Management Institute (PMI).** (2017). *A Guide to the Project Management Body of Knowledge (PMBOK® Guide), Sixth Edition*. Project Management Institute. 6 **Dudley, C.** (2024). The Rise of AI Governance: Unpacking ISO/IEC 42001. *Quality*, 63(8), 27-27. 7 **Dotan, R., Blili-Hamelin, B., Madhavan, R., Matthews, J., & Scarpino, J.** (2024). Evolving AI risk management: A maturity model based on the NIST AI risk management framework. *arXiv preprint arXiv:2401.15229*. 8 **Hansen, H. F., Lillesund, E., Mikalef, P., & Altwaijry, N.** (2024). Understanding artificial intelligence diffusion through an AI capability maturity model. *Information Systems Frontiers*, 1-17. 9 **Kubilay, B., & Celiktas, B.** (2025). Relationships Among Organizational-Level Maturities in Artificial Intelligence, Cybersecurity, and Digital Transformation: A Survey-Based Analysis. *IEEE Access*. 10 A**ven, T., & Renn, O.** (2010). *Risk management and governance: Concepts, guidelines and applications* (Vol. 16). Springer Science & Business Media. 11 **Sayles, J.** (2024). Aligning AI Governance with Other Internal Governance Models for Trustworthy AI:"The Convergence of Governance Frameworks". In *Principles of AI Governance and Model Risk Management: Master the Techniques for Ethical and Transparent AI Systems* (pp. 113-172). Berkeley, CA: Apress. 12 **Sayles, J.** (2024). Integrating AI Governance with Enterprise Governance Risk and Compliance. In *Principles of AI Governance and Model Risk Management: Master the Techniques for Ethical and Transparent AI Systems* (pp. 231-247). Berkeley, CA: Apress. 13 **Al Astal, A. Y. M., Ateeq, A., Milhem, M., & Shafie, D. I.** (2024). Corporate Governance and Internal Control Mechanisms: Developing a Strategic Framework. In *Business Sustainability with Artificial Intelligence (AI): Challenges and Opportunities: Volume 2* (pp. 551-564). Cham: Springer Nature Switzerland. 14 **Li, Y., & Goel, S.** (2025). Artificial intelligence auditability and auditor readiness for auditing artificial intelligence systems. *International Journal of Accounting Information Systems*, 56, 100739. 15 **Fernsel, L., Kalff, Y., & Simbeck, K.** (2024). Assessing the Auditability of AI-integrating Systems: A Framework and Learning Analytics Case Study. *arXiv preprint arXiv:2411.08906*. 16 **Sadiq, R. B., Safie, N., Abd Rahman, A. H., & Goudarzi, S.** (2021). Artificial intelligence maturity model: a systematic literature review. *PeerJ Computer Science*, 7, e661.

Index

T

U

V

W

X


ICTRECHT


ISBN 9789083567822