2024

Law & Compliance in AI Security & Data Protection

AI and Data Protection Training Module

MARCO ALMADA

Changelog

2024-12-17

First release of the course.

Contents

Introduction

Welcome to “Law & Compliance in AI Security & Data Protection”! This training module has been designed to support privacy and data protection professionals in their approach to artificial intelligence (AI). Over approximately 15 hours of self-study, the materials below will present an overview of the various stages of the life cycle of applications powered by AI technologies, from the initial stages of their development to the end of their operation. At each stage, the training materials will identify issues that AI introduces and amplifies, as well as potential responses to them. By studying those materials, professionals will be better positioned to understand whether and how their organizations can use AI in accordance with legal requirements for privacy and data protection.

To fully understand how AI matters for data protection and privacy, it is necessary to analyse what is unique about AI technologies and their production. This training module does not assume that learners have experience with the technical side of things. It introduces any technological concepts that are necessary for discussion and does so at an abstract level. The module’s goal is not to turn data protection and privacy professionals into computer scientists, but to ensure they have the concepts needed to understand the issues at hand and the vocabulary needed for effective communication with software developers and other technical actors.[1]

The course assumes that the reader is familiar with the general concepts of the GDPR. Basic concepts—such as the notions of “data subject” and “processing”—are taken for granted so that the course can focus on what changes with AI. Contrastingly, the module offers a more thorough revision of specialized topics, such as the rules on automated decision-making and regulation by design, with an emphasis on their AI dimension. Furthermore, the course will introduce learners to the interplay between the GDPR and the EU’s new regulation on AI technologies, the AI Act (Regulation (EU) 2024/1689). It will not offer an in-depth treatment of national requirements or industry-specific legal requirements. However, the conceptual tools developed throughout the module can also be applied to the study of such legal instruments and their implications for data protection.

Learning Outcomes

By the end of this training module, learners will be able to:

[1] Learners who are interested in a deeper dive into technical matters can consult the companion training module developed for ICT professionals: Enrico Glerean, *Elements of Secure AI Systems.*

- **Identify** the core technical features of artificial intelligence technologies and the various stages of their life cycle in an organization.
- **Map** the uses of AI systems within their organization and the actors involved in each use, with special emphasis on identifying data controllers and processors.
- **Take stock of** how these AI systems utilize and (potentially) generate personal data, and of the implications of that for compliance with data protection duties.
- **Assess** the implications of technical and organizational measures for data protection throughout the life cycle of an AI system; and
- **Distinguish** between the various kinds of mechanisms for evaluating AI systems (technical audits, impact assessments, certification schemes), identify when such evaluations are needed, and the techniques available to carry them out.

Module Structure

The training module you are about to start consists of three parts, each structured around a theme. The first part introduces the learners to basic concepts of AI and the issues they raise for data protection law. The second part discusses risks that take place at various stages of the *life cycle* of an AI-based tool, from the initial decision to make use of such a technology to the end of its operation. Finally, the third part offers an in-depth treatment of selected topics that are critical for organizations intending to use AI systems in accordance with the requirements of data protection law.

Within each part of the module

Each part of the module is divided into units. A module unit deals with a specific issue within the subject matter outlined by the part it belongs to. For example, Unit 13 (the fourth unit of Part III) deals with the data protection issues raised by the use of large language models. Module units are designed to demand at least an hour of self-study, to allow the learner to assimilate the concepts and get some familiarity with how to use the concepts in practice.

Within each learning unit

A unit of this module consists of an introduction, three sessions, and a conclusion. The introduction presents the general structure of the issue the unit covers. It also provides an overview of relevant topics not discussed in depth within the sessions. The sessions contain the bulk of the course's contents, focusing on topics that must be mastered for a comprehensive view of the unit's issue. Finally, a brief conclusion to each unit summarizes key points and highlights common trends between the individual sessions.

Coming back to the example of Unit 13, its issue is data protection and large language models. The introduction briefly discusses what is unique about those models, so as to warrant a full unit. The three sessions, in turn, analyse (1) the implications of the use of such models to data protection compliance; (2) safeguarding measures that can be

adopted during the design of those models; and (3) safeguarding measures that can be adopted when the model is used in a particular context. Finally, the conclusion highlights the main actionable points of those sessions.

Finally, each unit finishes with a list of references about the topic it covers. While some references are cited in the unit's text, citations have been reserved for passages where the text quotes from a specific text or discusses an argument or result published for a specific paper. The references section offers a more comprehensive listing of all the sources that guided the formulation of the unit. As such, learners should look at the listed materials if they want to dive more deeply into a particular issue, learn more about specific tools, or look for the answers to specific problems they face in practice.

The anatomy of a learning session

Every session of this learning module begins with an outline of its learning outcomes, that is, of the knowledges and skills the session will develop. After presenting this outline, the session follows moves on to presenting the theory behind that topic, with examples showing how the concepts emerge in practice. The exposition in each session is largely independent from the others, but references to previously covered topics will be present whenever they are needed.

The bulk of the module's content is, therefore, placed within individual sessions. However, each unit also has an introduction that situates the topics covered by its sessions, and a conclusion that articulates topics that cut across more than one session. Likewise, the introduction to a part defines the overall learning outcomes and context for its units, and the conclusion to a part articulates common trends and shared issues across units.

Tailoring the module to your needs

This training module allows learners to follow their own path to learning. If you follow this textbook from start to finish, you will acquire the basic concepts and tools that needed for identifying and addressing data protection issues related to AI technologies. However, not all learners have the same needs, and so this module is flexible enough to support different learning approaches.

By following a modular structure, this textbook allows learners to mix and match learning elements according to their needs. If a learner is already familiar with some topics covered by the module, they can skim through those sections and focus on whatever topics they have not mastered yet. If a learner has a particular interest in a specific topic, they can jump to the part, unit, or session, using the course's internal references to refresh other concepts as needed. And, if a learner wants to gain deeper knowledge in a particular topic, they can follow the module's references as a springboard for further learning.

The module is oriented towards self-learning sessions according to a learner's timer availability. Still, its modular structure lends itself to adaptation for a longer, instructor-led training. If an instructor has 30 minutes (or even an hour!) available for each session, they can dedicate the additional time to exercises and discussion between learners. However, those extensions are not essential for the learning experience, and self-study based on the materials provided below is a feasible means to develop the necessary knowledge and competences for dealing with the challenges of data protection in the age and AI.

Case Studies

This learning module supports data protection professionals as they deal with the impact of AI in their practice. Given that organizations use AI technologies for a variety of tasks and in many ways, it would not be feasible to cover all (or even the most common) applications in a single training module. Furthermore, as we shall see throughout the module, the data protection implications of AI relate closely to how AI technologies are used within an organization. Accordingly, this module focuses on providing general tools that are relevant for present and future applications, but learners will need to fill in the gaps of their specific contexts.

Nonetheless, the training module utilizes three hypothetical studies throughout its sessions. By dealing with those three cases, the module illustrates how various aspects of data protection law play with one another. Reliance on examples also shows how organizations in different contexts use AI in diverse ways, which cannot be treated in the same fashion but require instead attention to the particulars of the AI systems being used and their operational context. Session 1.3 of this training module details the examples.

About the Contributors

The first version of this training module was drafted by Marco Almada. As of December 2024, he is a postdoctoral researcher in Cyber Policy at the University of Luxembourg, working on the law and regulation of AI technologies. Marco has a PhD in law from the European University Institute, with a dissertation on technology-neutral regulation. Before that, he obtained bachelor's and master's degrees in both law and computing and worked as a data scientist and AI policy researcher.

The first version benefitted greatly from the comments and guidance of Konstantinos Limniotis, Georgia Panagopoulou, Spiros Papastergiou, and George Rousopoulos, and from the support of Amandine Jambert and Sixtine Crouzet. The author of the companion training module, Enrico Glerean, also offered valuable comments, especially regarding the technical passages of this text.

Part I: Fundamental Concepts

By the end of this part, learners will be able to:

- **define** artificial intelligence from a legal perspective and associate that definition with technical concepts;
- **distinguish** the various technical components that are articulated in the design and operation of AI system;
- **illustrate** various modes how an AI application might fail to work as expected; and
- **articulate** how safety and security issues create risks to data protection and other fundamental rights.

These days, AI technologies seem to be everywhere. They appear in personal tools such as the personal assistants in our smartphones, in business tasks such as automating human resources processes, in government practices such as tax fraud detection, and everything in between. With such widespread applications, AI is relevant to the work of data protection professionals in organizations of the most varied sectors and sizes.

Analyses of the impact of those technologies face various obstacles. It can be difficult in practice to figure out how those systems work, given their reliance on complex mathematical models and computer science techniques. Organizations might also struggle to pin down what kinds of personal data used within a system and the legal basis that authorizes the processing of that data. Often, organizations might not have clear answers even to more fundamental questions such as *how does the output of an AI system affects things in practice?* or even *is AI used at all here?* Answering these questions demands not only an understanding of what makes AI unique from a technical standpoint. It also of the legal and economic factors that restrict how organizations can obtain information about the AI systems they use and develop.

Part I of this course offers the conceptual foundations needed for such analyses. Over four units, it discusses key factors that must be considered for evaluating the data protection implications of AI technologies:

- **Unit 1** situates AI as a data protection issue, highlighting the risks and opportunities that AI technologies create for the protection of fundamental rights, as well as the legal framework that applies to them.
- **Unit 2** provides a bird’s-eye view of technical concepts related to AI, defining key concepts without going into technical details.

- **Unit 3** then provides a brief introduction to the cybersecurity dimension of AI, highlighting risks that are unique to those technologies.
- Finally, **Unit 4** discusses how AI technologies might produce undesirable effects even if they are adequate from a cybersecurity standpoint.

The knowledge covered in these units will then support the legal analyses discussed in the rest of the training module.

Unit 1. Introduction to Artificial Intelligence and Data Protection

By the end of this unit, learners will be able to:

- **illustrate** risks and opportunities of using AI in various contexts.
- **describe** the core features of the three case studies.
- **explain** how the new EU instruments on AI relate to data protection; and
- **indicate** the core elements of the AI Act's regulatory framework.

Why is AI relevant from a data protection standpoint? In part, this relevance comes from the fact that many AI applications have personal data in their inputs and/or outputs. For example, an AI system might use various pieces of information about an individual (input data) to make an inference (output) about whether they would be a suitable hire for a business. But, as Unit 2 of this training module will discuss in more depth, personal data also plays a more structural role in AI, when it is used in the training processes that take place when an AI system is developed. This is why, for instance, the Italian data protection authority opened proceedings against ChatGPT in 2023, requiring its provider (the US company OpenAI) to adopt corrective measures. Considering how widespread the use of AI technologies is, their dependence on personal data suggests that data protection professionals need to look closely at whether that data is processed in accordance with EU law.

This is not to say that the use of personal data in AI is inherently undesirable. After all, it has the potential to bring a variety of economic and social benefits. Those benefits can range from personal convenience (a good recommender system, for example, might save you the trouble of looking for a product you need to buy but keep forgetting about) to societal advantages, as the adoption of AI in public sector applications is often proposed as a way to deliver better public services. Considering these benefits, the use of even substantial amounts of personal data might be justifiable if it complies with the requirements of data protection law.

But the use of personal data in AI is not without risks. Because modern AI applications require significant amounts of personal data for their development and use, the accumulation of personal data gives margin to various risks that are well known by data protection professionals, such as those of misuse or data breaches. In addition, AI technologies create or amplify various risks, as illustrated by various scandals concerning discriminatory decision-making by algorithmic systems. The requirements and safeguards created by data protection law thus become particularly desirable when AI is involved.

The main purpose of this unit is to show how the EU's regulation of AI technologies interacts with data protection law. For that purpose, **Session 1.1** offers a general discussion of artificial intelligence, introducing the risks and opportunities associates with those technologies. **Session 1.2** discusses how the AI Act complements data protection law by addressing risks that are specific to AI technologies. Finally, **Session 1.3** introduces three hypothetical cases that illustrate how the AI Act and the GDPR both apply to different uses of AI in the public and private sectors. Those cases return throughout the module as a source of examples for the various concepts we cover.

Session 1.1. The risks and opportunities of artificial intelligence

> By the end of this session, learners will be able to **describe** why AI technologies have become more common in the last few years and **identify** some of the benefits and issues created by that diffusion.

AI technologies are becoming ubiquitous in modern society, shaping our routines and business environments in profound ways. For instance, facial recognition tools, used in border control and building access, streamline security checks but also carry significant privacy implications. Social networks leverage AI-powered recommender systems to predict and influence what content users see. Generative AI tools like ChatGPT can produce a wide range of content, from casual text to sophisticated audiovisual materials, demonstrating both the potential and the unpredictability of AI outputs. These examples suggest that AI is not a novel and futuristic concept, but rather something that is already deeply integrated into routine processes and high-stakes decisions in our lives.

Beyond these visible uses, AI has also become a part of our social infrastructures. Many businesses around the world now use AI-powered technologies to carry out various internal tasks. Human resources departments increasingly rely on AI tools to screen out candidate applications, especially as candidates themselves sometimes use AI to tailor their profiles. Strategic decision-making in large companies is guided by various forms of data analytics, such as those concerning market performance. Chatbots are used increasingly as a first channel of contact with consumers, which only interact with humans for more complex queries. Many of those uses of AI are also present in the public sector, as governmental organizations rely on AI-powered tools to carry out various facets of their work. This means that, in many countries, both the private and the public sectors depend much on their use of AI technologies.

The widespread adoption of AI is driven by multiple factors:

- **Advances in machine learning** and neural networks have enabled AI systems to perform tasks that were previously thought to be impossible or impractical.

- The **declining costs** of data processing and storage, along with the increased availability of computational power, make AI solutions accessible to more organizations than ever before.
- In addition, the **digitalization** of everyday activities has generated an abundance of data, creating both the need and the opportunity to leverage AI for analysis and decision-making.
- Organizations, whether in the private or public sector, are often motivated by the **competitive pressure** to innovate and the fear of falling behind, which can lead to rapid and sometimes poorly thought-out adoption of AI technologies.

These and other elements lead public and private organizations to adopt AI technologies for a variety of purposes.

The usefulness of AI for organizations depends on the tasks that one intends to automate and the available technical capabilities. **AI systems excel in certain tasks**, providing clear advantages in efficiency and scale. Language translation tools, for instance, have made it easier for people to communicate across linguistic barriers, enhancing both personal and professional interactions.

Even when AI does not outperform human capabilities, it can still offer **cost-effective solutions**. A good example is the use of generative AI in marketing campaigns. While the content it produces may not always be of the highest quality, it can generate large volumes of personalized messaging at a fraction of the cost of traditional methods.

In some cases, AI enables **activities that would be impossible without automation**, such as comprehensive audits of tax filings, which can help governments uncover patterns of fraud more effectively than manual inspections could.

Seen from a data protection angle, however**, the rapid proliferation of AI technologies is not without significant risks**. A major concern is the reliability of AI systems. Despite their impressive capabilities, AI tools can sometimes **fail to perform** as expected, leading to potentially grave consequences. For instance, emotion recognition technologies are often marketed as tools that can detect a person's feelings based on facial expressions or voice tone. Yet, the scientific basis for these claims is weak, and the algorithms frequently produce misleading results (Stark and Hutson 2022). The **complexity** of AI models can make it difficult to identify errors or biases in their predictions, leaving users and regulators blind to potential flaws until they cause real-world harm.

Another concern arises when AI is used for **inherently problematic or unlawful purposes**, regardless of how well the technology performs. For instance, an AI system designed to make hiring decisions may inadvertently exclude certain demographic groups if it has been trained on biased data, reinforcing existing inequalities in the job

market. In such cases, the effectiveness of the AI can amplify rather than mitigate harm, as it systematically executes a flawed process more efficiently than a human could. Similarly, AI-driven surveillance tools may enable extensive monitoring of individuals without their consent, raising serious ethical and legal questions about the right to privacy.

The reliance of AI technologies on large datasets can also create significant **privacy risks**. AI systems are often trained on vast amounts of personal information, sometimes collected without proper consent, or used in ways that individuals might not expect. This can lead to unintended consequences, such as exposing sensitive personal details or allowing for intrusive profiling. For example, an AI model used to predict consumer preferences might draw on data from social media, shopping history, or even biometric information, potentially leading to privacy violations if this data is mishandled or shared without adequate safeguards.

To address these risks and harness the benefits of AI responsibly, the European Union (EU) has embarked on **regulatory initiatives** aimed at balancing innovation with the protection of fundamental rights. As we have seen in the introduction to this unit, data protection law itself plays a vital role in this protective scheme. Because AI systems are often built on personal data and rely on it for their operation, **data protection obligations remain in force**, and thus help address some of those risks. In the following session, we will discuss another piece of legislation that contributes to AI governance in the EU: the Artificial Intelligence Act, which establishes additional factors that data protection professionals must consider in their work.

Session 1.2. The AI Act (Regulation (EU) 2024/1689)

> By the end of this session, learners will be able to **describe**, at a high level of abstraction, the core features of the AI Act (Regulation (EU) 2024/1689) and **compare** them with the treatment of risks in the GDPR.

The AI Act is a recent piece of legislation. It was proposed in response to various concerns about AI technologies that were voiced in society. Some of these, like the risks discussed in the previous session, are hypothetical concerns. Others, instead, reflect real-world harms related to AI technologies that are already in use. See, for example, the *SyRI* case in the Netherlands, in which the courts ruled that a risk scoring algorithm proposed by the government did not respect the right to a private life. To address those concerns, the EU lawmakers proposed a regulation that is very different from the GDPR, as it is based on the laws governing product safety rather than on data protection law.[1] Still, the reliance of AI technologies on data means that the AI Act

[1] On the structural differences between the AI Act and the GDPR, see Almada and Petit (2025).

affects how organizations must deal with their data protection obligations. In this session, we introduce the overall logic that guides the AI Act, before looking into its specific regulatory provisions in the rest of the training module.

A significant difference between the GDPR and the AI Act comes from their **object**, that is, from what those laws regulate in the first place. The GDPR is directed at the processing of personal data, that is, what one does with the data. The AI Act focuses instead on the technologies used to do that processing. It regulates AI systems, which it defines as a type of computer system that can do tasks such as generating content, recommendations, or even making decisions.[2] The Act also features some rules directed at AI models, which are the components that allow AI systems to carry out those tasks.[3] Because they regulate different things, those laws follow different approaches.

One should not, however, overestimate the differences between the GDPR and the AI Act. They both create **obligations to minimize the risks** created by their regulated objects:

- Article 25 GDPR obliges data controllers to adopt measures and safeguards to deal with risks to data protection principles, while Article 32 GDPR establishes an obligation to address risks to cybersecurity.
- In the AI Act, the providers of high-risk AI systems are required to adopt risk management measures (Article 9 AI Act), while the deployers of those systems must adopt their own approaches to deal with risks that appear in a specific application (Article 26 AI Act), such as the impact assessments that are required in some cases.

However, **risk assessment in the AI Act is considerably narrower than it is in the GDPR**.

Two factors contribute to the narrower assessment. The first one is that the obligations of providers of AI systems are mostly **limited to technical risks**. The actors regulated by the AI Act are expected to deal with risks that can be addressed through technical means or by providing technical information (see, e.g., Article 9(3) AI Act). In this regard, the GDPR goes further. It obliges regulated actors to adopt both technical measures—such as changes to the AI model powering an AI system—and organizational ones, such as limiting the number of persons that can operate an AI system. It follows from this that compliance with the AI Act's requirements for technical design might not be enough to meet what the GDPR demands.

[2] See the full definition in Article 3(1) AI Act.
[3] On the distinction between AI systems and models, see Session 2.1 of this training module.

The second limiting factor is that the AI Act establishes a **top-down risk assessment**. It does not apply a uniform set of rules to all AI systems and models. Instead, it separates those systems and models into different classes, each subject to its own legal framework. While the providers and deployers of AI systems are still obliged to identify and address the risks those systems create in practice, such an assessment takes place within the categories defined by the AI Act. Accordingly, it is necessary to examine the criteria the AI Act uses to assign systems and models to those categories.

Three different frameworks for AI systems

When it comes to AI systems, risk classification is based on the **purpose** for which a system was designed. The AI Act features a list of **prohibited AI practices**. That is, it is illegal to use an AI system for any of the applications listed in Article 5 AI Act. For example, one cannot use AI to materially distort the behaviour of a person (or group of persons) in a way that causes or is likely to cause harm to them or to others, such as manipulating them into a poor financial investment.[4] This is because the EU lawmaker has concluded that no measures can make AI systems safe enough to use in those contexts.

Within the lawful uses of AI, Article 6 AI Act singles out some applications of AI (listed in Annex I and III AI Act). Any system designed for use in such an application is a **high-risk AI system**, unless it is covered by one of the derogations in Article 6(3) AI Act. Whenever a system is classified as high risk, it becomes subject to a harmonized legal framework, which means that the rules that apply to them are the same throughout the European Union. Most of the AI Act is dedicated to setting up that legal framework, and some of these provisions will be analysed in this training module.

Finally, the AI Act does not establish a general framework for AI systems that are not high-risk or prohibited. It creates some obligations that are specific to certain applications. For example, the providers of AI systems that interact directly with natural persons must make sure that those persons can know they are interacting with an AI system (Article 50(1) AI Act). The AI Act also obliges the providers and deployers of AI systems, regardless of their risk level, to foster **AI literacy** among those dealing with the operation and use of AI systems on their behalf (Article 4 AI Act). Yet, for the most part, it considers that the **risks of systems outside the two categories addressed above are covered by existing laws**, such as the GDPR and sector-specific regulation at the EU and national levels.

Cumulative requirements for general-purpose AI models

By definition, the idea of regulating based on a specific purpose does not work for AI models that can be used for various purposes. To deal with those **general-purpose AI**

[4] Article 5(1)(a) AI Act.

models, the AI Act follows a cumulative approach. It establishes that the **providers of all general-purpose AI models** must comply with EU law on copyright and make some information about the model available to different types of stakeholders.[5] For example, providers of general-purpose models must supply information and documentation about a model to those who want to incorporate this model to their own AI systems.[6] The core idea behind those requirements is that they allow other actors to comply with their own legal requirements. Somebody using a general-purpose model to create their own AI system will need to have information to know how to use the model, and the general public is given the right to know about how the model is created.

Some general-purpose AI models with high-impact capabilities are classified as **general-purpose AI models with systemic risk**, and subject to additional requirements.[7] The notions of "high-impact capabilities" and "systemic risk" are both defined in Article 3 AI Act. However, the classification as a model with systemic risk is based not on the interpretation of these definitions but on the application of technical thresholds defined in Article 51 AI Act. For example, that article introduces a presumption that any general-purpose that has required more than 10^{25} floating-point operations for its training has systemic risk. Alternatively, the Commission has the power to designate a model as having systemic risk if its capabilities are somehow equivalent to that of systems meeting the relevant thresholds. For the most part, the AI Act treats systemic risk as something that can be quantitatively measured.

If a general-purpose AI model meets the criteria for systemic risk, its provider becomes subject to additional obligations. The provider must, among other things, mitigate the systemic risks created by the model's high-impact capabilities.[8]. By following those requirements, a provider is—at least in theory—addressing risks that could not be addressed by the downstream providers, that is, by those who use a general-purpose AI model to build a system. So, the rules on systemic risk are designed to promote trustworthy AI throughout the value chain of AI technologies.

Applying the AI Act

As a product safety law, the AI Act frames its obligations in terms of AI systems and models. Yet these objects are not the ones that must actually fulfil the obligations. This task falls primarily to two actors mentioned above: the **provider** of an AI system or model and its **deployer**. Articles 22–25 AI Act also stipulate obligations for other actors, such as importers, but the bulk of the Act concentrates on providers and deployers.

[5] Article 53 AI Act.
[6] Article 53(1)(b) AI Act.
[7] Article 55 AI Act.
[8] Article 55(2) AI Act.

To put it shortly, a provider is responsible for placing the AI system or model on the EU market, while a deployer uses an AI system for one or more purposes. The compliance of those two actors with the AI Act's requirements is overseen by **market surveillance authorities**. It is now time to briefly examine those definitions.

Providing AI systems and models

Under the AI Act, a **provider** is anybody—a natural person, a legal person, or any other entity—that either develops an AI system or general-purpose AI model.[9] One is also a provider if they place an AI system or model on the EU market or put into service under their own name. This is the case even if they did not develop the AI system in question. For example, if the *RandomCorp* corporation hires some developers to produce an AI system that will be sold under the *RandomCorp* brand, it becomes the provider of that system.

Additionally, **one becomes the provider of a high-risk AI system if they modify the system or its intended purpose**.[10] For example, suppose the online marketplace *SillyMarket* has a successful customer service chatbot it hired from a provider *RandomCorp*. Based on that success, somebody at *SillyMarket* has the idea of modifying the chatbot into a tool that mediates disputes between buyers and sellers. This new use is a high-risk application under Point 8(1), Annex III AI Act, which was not foreseen by *RandomCorp* as a potential use case for their chatbot. In this case, the AI Act stipulates that *SillyMarket*, not *RandomCorp*, is the one subject to the obligations for high-risk AI systems.

It is also useful to distinguish between the provider of an AI model and the **downstream providers** that incorporate the AI model into their own AI systems. The model provider might be subject to the obligations concerning general-purpose AI models, including those on systemic risk if applicable. But, if *RandomCorp* uses a model supplied by *ModelCorp* to create a high-risk AI system, *ModelCorp* is not in principle obliged to ensure that the system complies with the AI Act's rules on high-risk. *RandomCorp*, on the other hand, cannot avoid compliance with its obligations by blaming issues on *ModelCorp*'s model, even if it has little power to change to that model. This is why the AI Act obliges *ModelCorp* to make information about its model available to *RandomCorp*.

Deploying AI systems

A **deployer** of an AI system is anybody—again, regardless of legal form—that **uses an AI system under their own authority**.[11] For example, a sole trader that uses an AI system to optimize their operations would be the deployer of that system. So would a public sector organization that decides to use AI to automate internal processes. Any

[9] Article 3(3) AI Act.
[10] Article 25 AI Act.
[11] Article 3(4) AI Act.

deployer is subject to the AI literacy duty imposed by Article 4 AI Act: they must make sure that the people operating AI on their behalf know about the capacities, impacts, and limitations of an AI system. Deployers of high-risk AI systems are subject to additional duties, laid down in Articles 26 and 27 AI Act and examined in Part II of this training module.

As an exception to the classification above, Article 3(4) AI Act also stipulates that using an AI system in a personal non-professional activity does not count as deployment. This means that somebody who uses an AI tool to research information, or to tinker with their own photos, is not subject to the AI Act's obligations for deployers. They remain nonetheless covered by the requirements of other applicable laws, including the GDPR.

Enforcing legal requirements

The AI Act's requirements apply throughout the life cycle of AI systems and models. Providers and deployers must ensure compliance when an AI system (or model) is first placed on the market, put into service, or used. But they must also ensure ongoing conformity to the Act's requirements, which might require adjustments to a system or model. It might even be the case that a previously lawful AI system or model must be withdrawn from the EU market because it can no longer be sold or used in a safe way. Complying with the AI Act, just like with the GDPR, is an ongoing effort.

Before an AI system or model can enter the EU market, it must be in conformity with the AI Act's requirements. **In most cases, conformity is assessed by the providers themselves**, who draw up documentation to attest that the requirements are observed. There are some cases in which the AI Act requires third-party certification, such as for the biometric applications listed in Point 1 of Annex III AI Act[12] and for AI systems that are products (or components of products) that are themselves subject to third-party certification.[13] This means, for instance, that the provider of a credit scoring system does not need to rely on an external certification body. It might, however, pursue external certification to build legitimacy for their product.

Once an AI system is on the market, providers and deployers are obliged to carry out **post-market monitoring** of the AI system.[14] If they perceive that a system that is already on the market or in service can harm fundamental rights or other values protected by the AI Act, they must take appropriate measures. To ensure that is done, the AI Act's market surveillance mechanism empowers a series of **market surveillance authorities**.

[12] Article 43(1) AI Act.
[13] Article 43(3) AI Act.
[14] Articles 9 and 26(5) AI Act, respectively.

Each Member State must nominate at least one market surveillance authority.[15] A market surveillance authority is granted extensive powers to investigate AI systems that create risks to the values protected by the Act.[16] Based on those powers, it has the power to request that providers and deployers adopt corrective measures or even recall an AI system from the market.[17] A market surveillance authority can also issue fines and other sanctions in case of non-compliance with applicable requirements.[18]

The AI Act stipulates that market surveillance authorities must have the resources and infrastructure to carry out these tasks.[19] It leaves Member States mostly free to determine what authorities will carry out the role. However, it specifies that the **market surveillance authorities designated by other pieces of EU law are responsible for the AI systems within their scope**.[20] For example, financial regulators are responsible for the surveillance of AI systems used in regulated financial activities. Therefore, it is likely that each country will have more than one AI supervisory authority. In that case, each Member State must designate one of those authorities as the single contact point for the purposes of the Act.

In contrast with the rules for AI systems, the **rules for general-purpose AI models are enforced in a centralized fashion.** Enforcement powers are concentrated in the AI Office, which is a part of the European Commission.[21] It is this authority that is responsible for defining the technical thresholds for systemic risk and by ensuring that providers comply with the Act's requirements.

Given the overlap between data protection and the use of AI, some have suggested that data protection authorities are well-positioned to be involved in market surveillance. In fact, the AI Act designates the European Data Protection Supervisor as the surveillance authority for AI systems used by EU institutions, bodies, and agencies. It remains to be seen whether Member States will follow that lead. But, even if they do not, data protection authorities retain the power to enforce data protection law against these models.

[15] Article 70 AI Act.
[16] See, e.g., Article 74(13) AI Act.
[17] Article 79 AI Act.
[18] Article 99 AI Act.
[19] Article 70(3) AI Act.
[20] Article 74(3) AI Act.
[21] Article 64 AI Act.

Session 1.3. Three hypothetical case studies

By the end of this session, learners will be able to **describe** the general features of the three hypothetical cases used as sources of examples throughout the training module.

As we have seen in the previous sessions, AI technologies can be used in many contexts and for many reasons. This variety makes is a challenge for AI regulation. It makes more difficult for regulator to pin down risk levels, and to create obligations that are relevant for all systems with a certain level of risk. For those of us designing training modules on AI, it also means that examples must cover many cases. Because both the GDPR and the AI Act apply to a substantial number of AI systems and models,[22] there are many specificities that one must consider. Without engaging with the specifics of various contexts, an analysis might be too vague to be useful. However, one cannot cover all training cases within a single course, given the variety of sectors that would need to be covered.

To address this problem, this training module relies on three hypothetical case studies. Those cases are representative of many AI use contexts in the public and private sectors. In each case, AI systems and models are used for a variety of purposes, relying on different approaches to development, and based on distinct types of personal data. Therefore, the use of these cases as examples throughout the module will help illustrate the broad range of factors that need to be considered when assessing whether AI is being developed and used lawfully within an organization.

Case study 1: Artificial intelligence at the University of Nowhere

The **University of Nowhere (UNw)** is a large public university, which has thousands of undergraduate and postgraduate students in all areas of knowledge. Among its main research units is a well-known Law School and a small computer science that is among the best European centres on AI and technical security. Over the past decade, the university has more than doubled its number of students. However, cuts in public funding to education have meant that the university was unable to hire a comparable number of new professors and administrative staff. In this context, **UNw** is currently evaluating whether and how AI technologies might assist in its functions.

It is not hard to find examples of proposed uses of AI in education. Under current EU law, some of those applications are listed as high-risk use cases.[23] If **UNw** decides, for example, to use an algorithm to decide which students are most likely to thrive in its law school, the ensuing system would be classified as high risk. The outputs of this system

[22] Maybe even most of them.
[23] Point 3, Annex III AI Act.

might affect a potential student's likelihood of pursuing a law degree at **UNw**, or of continuing their studies once admitted. Hence, the AI Act would oblige the university to conform to various requirements before it can put such a system into service.

The high-risk classification, in this case, is based on the impact such a system might have on the outputs of the system might affect various fundamental rights of the students. Their right to good administration[24] might be affected as an automated system takes decisions about their future without giving them a chance to be heard. Biased decisions by an AI system might fall foul of the right to non-discrimination[25] if they are based on protected grounds such as ethnic or social origin, age, or political opinions. Those rights must be considered in the interpretation of the AI Act's provisions, as well as of other risk-based requirements, such as the data protection by design requirement from Article 25 GDPR.

Other applications of AI that might support **UNw**'s activities would not be classified as high-risk AI under the AI Act. For example, the university might decide to create a chatbot that can answer to common student requests such as the generation of diplomas and academic transcripts. In this case, the AI Act stipulates that the system must be designed in a way that allow individuals to know that they are interacting with an AI system.[26] It also requires **UNw** to educate its staff regarding the chatbot's capabilities.[27] But, for the most part, the main source of legal requirements here would be data protection law.

The specific contents of the requirements imposed on **UNw**'s use of AI will be examined in the various sessions under Parts II and III of this training module. Before any such analysis, however, it is important to clarify two aspects of this case study: where **UNw** gets data from and how it procures its AI systems.

Regarding personal data, **UNw** has access to considerable amounts of data about its students and academic and administrative staff. This data includes information presented at enrolment, student grades and sanctions, and the salaries of all its staff. It also has the technical means to acquire information from external sources, such as scraping the social network profiles of people who make their affiliation with **UNw** public. Lastly, the university might rely on external data providers ("data brokers") to acquire information that it cannot secure directly, such as information about potential hires or students. A data protection professional will therefore need to determine whether those various sources have been procured lawfully.

[24] Article 41 EU Charter of Fundamental Rights.
[25] Article 21 EU Charter of Fundamental Rights.
[26] Article 50(1) AI Act.
[27] Article 4 AI Act.

As for procurement, **UNw** has a strong computer science department and a large ICT team. This means it can afford to develop its own AI systems and models, as well as to fine-tune existing AI models for their own purposes. If they need (or decide) to hire AI systems and models, they must follow a public procurement procedure to do so. Therefore, there is a tendency to do things in-house, though, as discussed in Unit 13 of this course, this does not mean **UNw** is entirely independent from external providers.

Case study 2: AI in a small business

A few years ago, a couple decided to open their own business of smart toys. After much work and diligence, their startup **DigiToys** seems to finally be taking off. It now commercializes a small but growing range of interactive toys with educative purposes. By incorporating AI tools into dolls, puzzles, and other children's toys, they aim to help children above the age of three to cultivate a healthier relationship with the digital world. Within this proposal, the company is particularly interested in ensuring the good reputation and the legal conformity of its products.

DigiToys currently has approximately thirty workers. Its team includes a handful AI developers, who work in fine-tuning large language models for use within the toys. It also includes two teams of data scientists, who use AI tools for analysing data. As a result, the company is unlikely to develop general-purpose AI models of its own, let alone those with complex risks. But it has the capabilities to use those models for their own systems, including as components of their own products.

In particular, their use of AI systems within toys might raise obligations under the GDPR and the AI Act. If the toys process personal data, they become subject to EU data protection law. Furthermore, the company's concern with safety means that it has opted to follow a third-party certification procedure for its toys.[28] As such, its toys are covered by Article 6(1) AI Act, and therefore subject to the rules on high-risk AI.

Additionally, **DigiToys**'s data scientists also make use of AI systems. Their product team uses AI to analyse large volumes of data about the toys, which stem from sources such as consumer satisfaction reports as well as telemetric data and error reports from each individual toy. These analyses are used to diagnose errors in toys, identify if they are having a healthy effect on the behaviour of children, and to produce ideas for new products. None of those applications is covered by the list of high-risk AI applications in Annex III AI Act. Still, the data used for those analyses is likely to contain significant amounts of personal data from interaction with children.

Data scientists in **DigiToys**'s marketing team rely on data from other sources. In fact, the company goes to a great length to make sure marketing never has access to data collected from products. Marketing operations rely instead on information sourced from

[28] See Article 19 of Directive 2009/48/EC, which harmonizes the rules on toys in the EU.

the company's customer databases and from online advertisement platforms. That information is used to segment potential and actual customers into profitability groups, as well as to offer personalized product recommendations to them. Once again, those applications fall outside the high-risk classification in the AI Act, but they involve substantial volumes of personal data about the adults that buy (or might buy) toys for their children.

Case study 3: Data-driven medical technologies

The hospital **InnovaHospital** is a private, non-profit medical organization that has branches all over the country. Over the past few decades, it has acquired a reputation for rigorous observance of patient confidentiality and data protection requirements, particularly for its serious response in the few times data leaks and other breaches took place. It is also known for its openness to innovation, as it hires healthcare professionals that are always working on the development of new techniques.

Within **InnovaHospital**, executives have identified two priority areas for the application of AI technologies. First, they want to use AI technologies to streamline their human resources department, spotting talent and helping its development from early on. This application would be classified as high-risk under the AI Act,[29] as it has the potential to affect the careers of everybody hired by the hospital and, in doing so, affect their rights as an employee. To create such a system, the hospital has access to its internal data keeping, such as evaluation reports, as well as data it collects during the hiring process. Some decision-makers have also considered acquiring data from additional sources, such as the social networks of new hires.

Second, they want to evaluate whether and how they can use patient data to develop technologies that support clinical practice. As examples of the ideas that have been raised include, one can see the use of data from patient exams to train AI systems that can be used as medical devices[30] or for personalizing the treatment given to each patient.

One obstacle that **InnovaHospital** faces in its use of AI is that, despite its large availability of data, it does not have the ICT capabilities needed to develop cutting-edge AI technologies on its own. As such, it will need to hire new professionals, buy ready-made AI solutions, or rely on AI-as-a-service solutions purchased from a provider. Each of these solutions has its own drawbacks, which will come up at various points in this module.

[29] Point 4, Annex III AI Act.
[30] Which means that in some cases they are covered by Article 6(1) AI Act.

Conclusion to Unit 1

AI technologies can take many forms, and they can play many roles within organizations. In many of these roles, the creation and use of AI systems and models is highly dependent on personal data. As such, data protection law is an important piece of AI governance, and the AI Act does not make the GDPR redundant. If anything, the latter becomes **more** relevant, both because of direct mentions and because AI regulation creates better conditions for applying data protection law for AI technologies. Still, it is undeniable that the result is a complex legal framework, even for seasoned data protection professionals.

This unit has supplied an overview of the AI Act's regulatory framework. Such an overview is necessarily abstract, given that the Act covers a vast range of applications which cannot all be treated in the same way. Just like the GDPR, the legal requirements remain the same, but the risks that need to be tackled in each context can be vastly different from one another. By understanding the overall logic behind the Act, you will now be better positioned to understand how its requirements interact with the GDPR. This knowledge will provide a starting point for the rest of the module. Therefore, take your time to revisit this session before moving forward. Doing so will pay off in the longer run.

Prompt for reflection

Discuss how the AI Act's classification of risks (prohibited, high-risk, and other AI systems) helps balance innovation and fundamental rights. Consider whether this approach is sufficient to address emerging AI challenges and whether it complements the GDPR effectively.

References

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) [2024] OJ L.

European Commission, 'Proposal for a Regulation of the European Parliament and of the Council Laying down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts' (COM(2021) 206 final, 21 April 2021).

Marco Almada and Nicolas Petit, 'The EU AI Act: Between the Rock of Product Safety and the Hard Place of Fundamental Rights' (2025) 62 Common Market Law Review.

David Fernández-Llorca and others, 'An Interdisciplinary Account of the Terminological Choices by EU Policymakers Ahead of the Final Agreement on the AI Act: AI System, General Purpose AI System, Foundation Model, and Generative AI' [2024] Artificial Intelligence and Law.

Gabriele Mazzini and Salvatore Scalzo, 'The Proposal for the Artificial Intelligence Act: Considerations around Some Key Concepts' in Carmelita Camardi (ed), *La via europea per l'Intelligenza artificiale* (Cedam 2022).

Claudio Novelli and others, 'AI Risk Assessment: A Scenario-Based, Proportional Methodology for the AI Act' (2024) 3 Digital Society 13.

Kasia Söderlund and Stefan Larsson, 'Enforcement Design Patterns in EU Law: An Analysis of the AI Act' (2024) 3 Digital Society 41.

Luke Stark and Jevan Hutson, 'Physiognomic Artificial Intelligence' (2022) 32 Fordham Intellectual Property, Media and Entertainment Law Journal 922.

Sandra Wachter, 'Limitations and Loopholes in the E.U. AI Act and AI Liability Directives: What This Means for the European Union, the United States, and Beyond' (2024) 26 Yale Journal of Law & Technology 671.

Unit 2. Core Concepts of Artificial Intelligence

By the end of this unit, learners will be able to **discuss** technical concepts of AI and **explain** to technical stakeholders how those concepts are relevant to data protection debates.

In the previous unit, we discussed how AI creates risks and opportunities that are relevant for compliance with data protection obligations. To understand *how* AI does so, one must understand how AI-powered technologies work. This is what we will do in this unit of the course.

The first thing in that discussion is to examine what we are talking about when we talk about “artificial intelligence.” For some people, AI conjures ideas of helpful technologies, such as the personal assistants in our smartphones. For others, it creates apocalyptic ideas out of science fiction, such as robots rebelling against their human masters. But AI can also make people think about very real risks, such as those covered above. So, the term can mean different things for different people, and those impressions are often coloured by fiction and by individual experiences. A clear discussion of the impacts of AI requires common ground for debate.

For the purposes of our training module, when we talk about “AI” we are talking about a **technical practice**. That is, “artificial intelligence” is what computer scientists, statisticians, and other technically people do when they want to solve certain technical problems. For example, if one wants to create a recommender system, they can use various approaches to do so, such as creating a machine learning model based on consumption habits of the users of a platform. Under this definition, it makes no sense to say that “an AI” did something, because AI is an abstraction.

It follows from this definition that an analysis of the legal relevance of AI should be more specific. It should name the techniques and the technical objects that are of interest because different technical choices can have different impacts in the real world. For example, creating an AI system based on machine learning technologies requires a considerable amount of data, but it can lead to the successful performance of tasks that were not feasible with previous expert systems. This unit provides some of the terms that data protection professionals need to know in order to make relevant distinctions.

Our goal for the following three sessions is not to turn data protection professionals into technical experts. Because of the complexity of AI technologies, developing such a competence would require time and effort that are not reasonable to expect from data protection professionals that are already overloaded. In fact, a narrow technical introduction to AI concepts (such as an introductory concept) can be misleading, as it might obscure complexities that appear in the real world. Furthermore, the specifics

might soon become outdated as technology evolves. Instead, this unit offers an introduction to **basic concepts of the technological side of AI**.

Those concepts can play two roles. First, they can help in **exercising critical reasoning** regarding technologies. As we shall see in Unit 4 of this course, AI technologies (as any other technologies) do not always live up to what they promise, and knowing where to look can help us not to be swindled by sales pitches. Second, a good grasp of the terminology can be useful for **dialoguing with technical experts** within an organization, as well as with contractors. As such, the basis offered by this unit should remain useful in practice even after the technologies that are the state of the art today are retired.

For that end, this unit focuses on three aspects of AI technologies. **Session 2.1** looks under the hood of AI technologies and defines the procedures that are used to create them. **Session 2.2** then discusses the relationship between data and artificial intelligence, while **Session 2.3** concludes the unit by discussing the technical infrastructures that allow all that to function.

Session 2.1. How AI works

> By the end of this session, learners will manage to **distinguish** between the main technical approaches used to build AI systems and **identify** the core features of each approach.

The logic that guides AI technologies can sometimes seem arcane. For example, the internet is full of examples where a chatbot is fooled into giving a silly answer to a question because that question is phrased in a peculiar way. However, the details of those AI technologies shape how they work and produce effects in practice. As such, a good understanding of them is essential for properly applying the relevant law to their design and use. To support this understanding, we will begin by the core of what makes AI unique — its algorithms and models.

At its essence, an AI system is a type of computer program, executed by a computer in the same way as any other software. Like all computer programs, AI technologies rely on **algorithms**. An algorithm is simply a set of step-by-step instructions that tell the computer how to solve a problem or perform a specific task. You might think of it like a recipe. Given certain ingredients (input data), the algorithm tells you what steps to take to prepare a dish (the output). A familiar example is the long division algorithm, which provides a series of steps to divide one number by another, producing both a quotient and a remainder.

In the context of AI, the term “algorithm” is often used to refer to the entire decision-making process of the AI system. For instance, someone might say, “The algorithm recommended this video to me,” even though the result is actually produced by a complex set of algorithms together within a platform rather than by a single procedure. This kind of shorthand reflects the leading role algorithms play in AI technologies, as they define the rules and logic that produce the system’s outputs.

A huge portion of modern AI systems relies on machine learning, a type of AI technique where the specific algorithm for producing outputs is not manually programmed by a developer. Instead, the algorithm that generates the outputs is itself configured by a **learning algorithm** that processes enormous amounts of data to learn patterns that can be generalized for future decisions. Although there are other approaches to AI, such as expert systems that rely on pre-defined rules, machine learning has been the dominant force behind recent AI advancements. As such, we will focus our discussion on them.

Machine learning approaches

The term **machine learning** refers to a broad family of ways to create AI systems. For the purposes of this training module, it is important to distinguish between three main classes of approaches:

1. **Supervised Learning** is the most common type of machine learning. In supervised learning, the algorithm is trained using a labelled dataset, which means that the input data comes with corresponding correct outputs (labels). The system learns by comparing its predictions to the correct answers and adjusting its internal model to reduce errors over time. For example, a supervised learning algorithm might be trained to recognize cats in photos by being shown thousands of images labelled “cat” or “not cat.” Through this process, the system learns to generalize from these examples and can eventually identify whether a new, unlabelled photo contains a cat.

2. **Unsupervised Learning** involves training an algorithm on data without any labelled responses. Instead of learning from examples, the algorithm tries to find patterns or structures within the data itself. One common use of unsupervised learning is in clustering, where the algorithm groups similar data points together. For instance, a company might use unsupervised learning to segment customers into distinct groups based on their purchasing behaviour, even if the system was not told what kinds of groups to look for.

3. **Reinforcement Learning** trains algorithms through their interaction with a physical or virtual environment. As it interacts with that environment, the algorithm receives feedback on its actions, allowing it to learn from trial and error. Successful actions lead to rewards, while mistakes lead to penalties. An example

of reinforcement learning is training an AI to play a video game: the algorithm tries different strategies, learns from the rewards (such as points scored in the game), and improves its play over time.

The result of this learning process is an **AI model**, which is a representation of what the system has learned from the data. The model contains the decision-making logic that the AI system uses when processing new inputs. One common type of AI model is the **neural network**, which is inspired by the structure of the human brain. A neural network is made up of layers of artificial neurons, each of which performs a simple computation. The neurons are organized in layers, and the output from one layer serves as the input to the next. During training, the neural network adjusts the connections between neurons to improve its performance on the given task.

Let us break down how a neural network works in practice. Imagine a system designed to recognize handwritten digits. When you provide an image of a handwritten number, the neural network processes the image through multiple layers of neurons. Each neuron combines the input data in a specific way, applying weights and biases that were adjusted during the training phase. The final layer of the network produces an output, such as predicting which digit (0-9) the image represents. This output is based on the rules and patterns the model learned during training.

It is important to understand that a neural network, like other AI models, does not "know" the answer in the way a human does. Instead, it applies complex mathematical transformations to the input data based on patterns it has seen before. This means that while neural networks can be highly effective, they can also be opaque or difficult to interpret, a phenomenon often referred to as the "black box" problem, as we will discuss in Session 4.3 of this training module.

From models to systems

An AI model is an object that can be used to perform the task(s) for which it was trained. Many models are created for a specific purpose: the sample neural network described above can only recognized tasks, and one would have to train an entirely new model to recognize dogs. In recent years, however, there is a growing number of **general-purpose AI models**, which are trained for a variety of tasks. For example, OpenAI's GPT family of language models can generate several types of content, such as conversations in which they interact with humans or large texts about many subjects. These models are sometimes called **foundation models**, as they work as a building block for many types of AI systems.

So, **what distinguishes an AI model from an AI system?** Sometimes, the terms are used interchangeably. Yet, the AI Act distinguishes between them, as do some technical sources. Following this distinction, the AI model is a component that allow the AI system to carry out the tasks that we think of as "artificial intelligence" tasks. For

example, a recommender model allows a social media platform to suggest posts to a user based on that user's previous interactions with content. An AI system (at least one based in machine learning) will include an AI model, but it will also feature other components. So, the difference between them is akin to the difference between an engine and a complete car.

To get from an AI model to an AI system, one needs to add various kinds of components:

1. To operate, an AI model needs access to input data, which might be collected from various sources or provided by user interactions. For example, a recommendation system in an online platform might draw on user preferences and browsing history to suggest new content, and that information is collected by tools such as cookies.
2. Once a model operates, its outputs need to be delivered *somewhere*. This can be a database where records are stored, a chatbot interface, or a dashboard displaying predictions or recommendations, among other possibilities.
3. An application might interact with the AI model through an **API (application programming interface)**. An API is a set of rules and protocols that allows different software applications to communicate with each other. It acts like a bridge, enabling one program to request data or services from another without needing to understand the internal workings of the other system. For example, many of the applications powered by large language models such as GPT-4o do not replicate those models in the application itself but communicate with a centralized model through an API.
4. As we shall see later in this training module, effective AI systems include monitoring tools to track performance and detect any issues that might arise in real-world use, such as shifts in data quality or unexpected model behaviour.

Those are just some examples of components that can have an impact on how a system functions. Even if they are not powered by AI techniques themselves, they can affect the impact an AI system has in the world. As such, they become directly relevant when one is assessing that system's compliance with legal requirements.

In short, AI systems are driven by algorithms, with machine learning algorithms playing a dominant role in recent advancements. These systems learn from data, creating models that represent patterns and relationships. While this approach offers powerful capabilities, it also comes with challenges, particularly in terms of transparency, data privacy, and potential biases. By understanding the basic concepts of AI algorithms, data protection professionals can better navigate the complexities of AI technologies and advocate for practices that protect individuals' rights.

Session 2.2. Personal data in AI systems

By the end of this session, learners will be able to **identify** the various roles personal data plays in AI systems: as inputs for the training process, as inputs for their use, and as outputs of the system's operation.

As a technology-neutral regulation, the GDPR largely refrains from distinguishing processing in the training process from other kinds of processing. Yet, the specific uses of AI data in the creation and use of AI systems and models raises some concerns that are not present in other types of data processing, or at least are not as salient there. For example, the large volumes of personal data used to create high-end AI models can lead to massive privacy breaches if that data somehow leak. Those issues coexist with more general issues, such as the need to find a legal basis for the processing of any personal data used in this context. This session supports data protection professionals by offering a brief introduction to how personal data can come into play in AI.

To put it shortly, personal data can play three roles when it comes to AI systems:

1. Personal data can be an **input** to the operation of an AI system. For example, a recommender system might take information about the personal interests of a user in a social media platform to find out what content that user would like to see.
2. Personal data can also be the **output** of the operation of an AI system. For example, an AI system created for creating risk scores for a crime (such as financial fraud) receives information about an individual and then ascribes to that individual a risk score that represents their likelihood of committing that crime.
3. Personal data can be a **building block** for an AI system or model. For example, a machine learning model that is intended for the kinds of tasks above will likely be trained on data about individuals that are relevant for the problem, such as platform users and previous investigations of financial fraud, respectively.

As the examples suggest, those uses are often interconnected. A system that is meant to process personal data will likely generate outputs that can be associated with individuals,[1] and personal data will be used in its construction process to ensure the quality of its outputs. In this session, we will look at the various approaches organizations can use to obtain data for their AI systems. Before that, however, we will briefly discuss the roles data can play in the construction of an AI system.

[1] Though not always. The output might, for example, be a statistical aggregate of individual properties that cannot be traced to a single individual.

The AI Act, following technical practices, distinguishes between three types of data sets that are relevant in the construction of an AI system:

- **Training data** refers to the data to which the learning algorithm is applied,[2] that is, to the data from which the patterns contained in the finished model are generated.
    - In the case of a supervised learning model, this will usually be a set of examples that pair some input data with the expected output.
    - For unsupervised learning models, no expected outputs are provided, just the input data.
    - For reinforced learning, one does not supply expected responses, but the system must be given information about the payoff of different options.
- **Validation data** is used for tuning the trained model, allowing the model builders to choose between different learning processes and strategies.[3] For example, it allows builders to avoid the phenomenon of overfitting, in which a model learns rules that describe well the training set but do not generalize well.
- **Testing data** is used for evaluating the overall performance of the AI system before it can be sold or placed into service.[4] That is, it provides a base for evaluating the system *after* any technical validation processes.

For AI systems that are not built from machine learning techniques, testing data will still be necessary to evaluate their performance in the intended test cases. If one or more of those datasets contains personal data, data protection law is likely applicable to their processing. And, since the learning process and the comparison of test data with model outputs both require processing, this means data protection becomes relevant for the training process, too. Hence, we will now consider how organizations might secure data for their needs as they build and use AI.

Directly collecting data

An organization can start measuring some kinds of data that are relevant for the application they want to develop. That data can take various forms, such as:

a. **Measuring user interactions**: For example, **DigiToys** might collect data on how often children interact with their toys, or on their speech patterns, for the design of product updates.
b. *Analysing internal data*: For example, the **UNw** can use its raw data about students to generate metrics, which might later be fed into an AI system.

[2] Article 3(29) AI Act.
[3] Article 3(30) AI Act.
[4] Article 3(32) AI Act.

c. *Creating new data from the combination of existing sources*: For example, **InnovaHospital** might integrate patient data from different branches of its operations to obtain a holistic view of patient health.

When it collects that data, the organization becomes a data controller for the operations involved in collecting this data and directing it towards AI.

Reutilizing personal data

Some organizations amass personal data as part of their operation. For example, a hospital cannot carry out its core functions without information about its patients. That data might be an asset for the development of AI technologies, but its use is subject to legal constraints that are discussed later in this section.

A few data quality issues might reduce the usefulness of previously available data:

1. **Relevance**: one needs to evaluate whether the dimensions captured in existing data are relevant for the problem the AI system or model is meant to solve. For example, the **UNw** university might use data about the courses each student follows to schedule its purchase of library books, but the that data might not be particularly useful for creating a chatbot.
2. **Assumptions embedded in data**: despite what the term “raw data” might suggest, even the most comprehensive datasets contain some assumptions in them: what data is relevant enough to be stored, how should this variable be measured, how to treat missing values, and so on. If unchecked, those assumptions can create problems. For example, if **InnovaHospital** wants to create a tool for supporting the diagnosis of heart attacks, that tool must account for the differences in symptoms between men and women. Otherwise, it might focus on the metrics that usually reflect male symptoms and fail to serve more than half of the population.
3. **Errors, outdated data, and missing data**: one must be aware of what issues are present in the existing dataset and how they are managed. For example, how does **DigiToys** treat duplicated information received from toys? What error correction mechanisms does it adopt on the transmitted data?

Acquiring data from third-party brokers

Many organizations (the so-called “data brokers”) have a business model that is based on the commercialization of data about individuals and organizations. If an organization decides to acquire data from them, it should exercise caution. The same data quality issues outlined above remain relevant here.

Additionally, one must consider whether the broker has lawfully obtained control of that data and whether there are legal bases for the transfer. Some models of brokerage have already been questioned from a legal perspective, leading to some enforcement

decisions and ongoing cases. Hence, an organization needs to exercise due diligence when procuring data for third parties and consider how their AI system or model will be impacted if that business model is found to not comply with the GDPR.

Building synthetic data

Sometimes, an organization cannot rely on fully anonymized data. If an application involves the profiling of natural persons, for instance, it cannot be trained or used without some form of reference to such a person. For example, an AI system for medical diagnoses will eventually be used in someone, generating a piece of personal data about them (their health status). Given that the use of large-scale personal data for such applications can be risky, some organizations have proposed the use of synthetic data as an alternative.[5]

Because synthetic data does not refer to an actual person (identified or identifiable), it would fall outside the GDPR's definition of personal data. So, to the extent that the synthetic data offers a faithful reproduction of the population to which the AI system applies, it would allow the use of AI without creating data protection risks.

The exemption from data protection law only applies if the data is *actually* synthetic. If it is possible to find information about natural persons based on the synthetic dataset, it remains covered by data protection law. This is the case even if the values ascribed to that person do not match reality. For example, consider a situation in which a synthetic database keeps the real names of people for credit scoring, but assigns them random values for each metric. That database will not allow an observer to discover correct information about the named individuals. Still, it associates that information to their identities, and the GDPR's definition of personal data features no exception for incorrect information.

Even if the data itself has no association with an identified or identifiable natural person, data protection law might also apply to its generation. This is the case if the synthetic data is generated from a dataset containing information about actual natural persons. While the ensuing database might not be personal data, creating it requires the processing of personal data. For example, **InnovaHospital** might use create a synthetic dataset from some of its medical records. In that case, the hospital remains obliged to the follow the GDPR as it creates the dataset, though the use of that dataset might not be covered by it.

Regardless of its legal classification, synthetic data remains subject to the data quality issues raised above. This kind of data is not a silver bullet for the construction of AI. Still, it can be useful if used judiciously.

[5] On the concept of synthetic data, see Session 2.2.

Session 2.3. The technical infrastructure of AI

By the end of this session, learners will be able to **distinguish** between the various components of the “stack” that supports the execution of an AI system.

While discussions about Artificial Intelligence (AI) often focus on algorithms, models, and data, it is essential to understand that these are all abstractions — simplified representations of what is happening under the hood. Ultimately, an AI system is a computer program, which relies on the underlying technical infrastructure to function. This infrastructure includes not just the computers executing the code but also the networks and storage systems that provide the necessary resources. In this session, we will introduce the main elements of this infrastructure and discuss how they can matter for data protection purposes.

Computing power as a need for AI

Let us start with the concept of **compute**. In technical terms, compute refers to the processing power required to run an AI program. Compute power is what allows an AI system to process data, execute complex algorithms, and generate outputs.

While a typical laptop might be sufficient for running simple AI tasks, the training process of more sophisticated AI models — such as those used in natural language processing or image recognition — requires much greater compute power. Depending on the scale of the model, even running a model that might be already trained can demand many resources. These tasks often rely on specialized hardware like Graphics Processing Units (GPUs) or Tensor Processing Units (TPUs), which are designed to handle the heavy computational loads involved in AI training and inference.

One measure that is often used to capture how much compute is used is that of **floating-point operations** (FLOPs). Without going into much technical detail, a FLOP is a type of mathematical operation that happens within a computer processor. Training a large AI model requires a substantial number of these operations. For example, the rules on systemic risk under the AI Act apply (by presumption) to advanced models trained over more than 10^{25} FLOPs, that is, more than ten septillions of those mathematical operations. A few of the models that exist nowadays, such as Google’s Gemini or OpenAI’s GPT-4o, are said to exceed this threshold.

As of 2024, most of the compute costs in AI training happen during the training process. However, as some studies suggest (Erdil 2024), there is a **trade-off between compute during training and compute at inference time**, that is, at the moment when an AI system is expected to generate its outputs. There are strategies that allow model

builders to reduce the costs involved in training, but at the expense of increasing the number of operations that a trained AI system must perform to generate output.

This trade-off can have implications for organizations using pre-trained AI systems. Each FLOP a processor executes costs a tiny bit of energy and takes some time. The amounts for each operation are vanishingly small, but, as we have seen, there are *many* operations involved even in the simplest AI tasks. This means that a model that does its most to reduce compute costs at inference time can be cheaper to use, even if at a greater expense to its creator. Conversely, developers might reduce their training costs in a way that makes it more expensive to run the finished AI system.

Memory and storage of data in AI systems

Compute is not the only physical factor at play when it comes to AI systems. Those systems rely heavily on memory and storage, that is, on physical supports that allow a computer to store and process information. The information that needs to be preserved includes not just the system's output and its input, but the intermediary steps involved in the enormous number of calculations described above. As a result, both the training and use of AI systems can be dependent on the availability of means for memory and storage.

Memory is used for temporarily holding data that the AI system needs to access quickly while processing tasks. The more memory available, the more data the system can access while executing its model. However, memory is volatile — it only holds data temporarily. Once a program finishes its execution, it will ideally free up memory for the next one. For example, the memory used to make an inference about a user's preferences for a recommender system will likely be overwritten when the system makes a reference for another user.

Sometimes, a computer needs to preserve information for longer. For example, when one generates data as the result of an AI system's operation, there is usually some interest in preserving that data. To do so, computers rely on **long-term storage**, such as hard drives or solid-state drives. Those sources of storage can retain information for a long time, without requiring the kind of active effort needed to preserve memory. The trade-off, here, is that reading information in long-term storage is much slower than reading information in memory. In fact, one of the major sources of delay when a program is running can be the time that is spent taking information from long-term storage and sending it to memory when it needs to be used often. But, since storage devices are cheaper and more lasting than memory, they are essential for storing large datasets and pre-trained AI models that can be used repeatedly, as well as the data one needs to preserve.

Network connectivity

Many AI applications are dependent on the flow of information from other devices. For example, AI systems used in social networks rely on the internet to transmit and receive information. This means that the properties of internet connection, such as download speed and bandwidth, become particularly relevant for their operation.

For instance, a virtual assistant on a smartphone might need to send a voice recording to a cloud server for analysis, requiring a fast and reliable internet connection. If the network speed is insufficient, the response time might lag. If that happens, the user's experience is negatively impacted, even if the AI system manages to generate inferences quickly enough.

The cloud as an AI enabler

As discussed above, running anything but the most trivial AI systems requires a lot of resources. However, few organizations have the financial wherewithal or the technical capabilities to maintain all that technical infrastructure. Therefore, the use of AI models and systems has been incredibly facilitated by the fact that individuals and organizations can contract the use of those resources through cloud platforms.

A **cloud** is a network of remote servers that provide computing power, storage, and other resources over the internet. These resources are made available for customers, who can, for example, acquire access to a machine by paying a fee based on time or on the amount of resources used. When an AI application is described as "cloud-based," it means that the heavy computational tasks are not performed on the user's device, such as a smartphone or laptop. Instead, the heavy work of processing data and making AI inferences is carried out on powerful servers maintained by cloud providers. This setup allows organizations to access vast amounts of computing power without investing in expensive hardware, making it easier and more cost-effective to deploy AI technologies.

However, cloud computing also raises important considerations for data protection. Storing data in the cloud means outsourcing the maintenance and security of that data to a third party. While this can offer benefits in terms of scalability and cost, it also introduces potential risks.

Many major cloud providers, such as Amazon Web Services (AWS), Microsoft Azure, and Google Cloud, are based outside Europe. This raises concerns about cross-border data transfers and compliance with the GDPR, as seen in the general concerns about cross-border data transfers. Organizations will need to cope with other potential sources of risk as well, such as potential vulnerabilities in the cloud infrastructure that could be exploited by malicious actors.

Finally, cloud platforms have a variety of reliability mechanisms. Nonetheless, they are still a single point of failure outside the control of the organization relying on them. It

follows from this that a cloud outage can reduce the availability of many services at the same time, as all services relying on a given provider will be affected by its failures. Organizations need to take these potential risks into account when considering the savings and other advantages they might derive from relying on a cloud provider.

Conclusion to Unit 2

Why does a training module with a legal focus need to zoom into the technicalities of AI? After all, the GDPR is designed to be a technology-neutral regulation,[6] which means its provisions apply regardless of whether data is processed by AI or another technological arrangement. Even so, there are several reasons why technical understanding can be helpful for data protection professionals.

Sometimes it is possible to adequately describe problems with “algorithms” and “models” without going into technical details. For example, one can identify algorithmic biases by looking at the outputs of AI systems rather than inspecting their inner workings.[7] This means that abstractions can help us make sense of why AI matters from a legal reason. However, abstractions in computing are always “leaky,” in the sense that the technical details that are abstracted away can sometimes have significant real-world implications.

For example, a defect in the processor used by a cloud server could lead to errors in the AI system’s calculations, producing incorrect or biased results (see, for example, Hochschild et al. 2021). Similarly, a security vulnerability in the cloud provider’s infrastructure could allow unauthorized access to the data being used by the AI system, potentially compromising sensitive personal information.

Given these risks, data protection professionals need to take a proactive role in assessing the technical infrastructure of AI systems used by their organizations. This includes evaluating the security measures implemented by cloud providers, understanding where and how data is stored and processed, and ensuring that cross-border data transfers comply with relevant legal requirements. By gaining a basic understanding of the infrastructure that supports AI, data protection professionals can better identify potential vulnerabilities and work towards mitigating risks.

Prompt for reflection

Reflect on the distinction between an AI system and an AI model. Why is it important for data protection officers to understand this distinction when evaluating compliance with legal requirements?

[6] Recital 15 GDPR.
[7] On bias, see Session 4.2 of this training module.

References

Marianne Bellotti, *Kill It with Fire: Manage Aging Computer Systems* (No Starch Press 2021).

Kate Crawford, *The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence* (Yale University Press 2021).

Ege Erdil, Optimally Allocating Compute Between Inference and Training (Epoch AI 2024).

Peter H Hochschild and others, 'Cores That Don't Count', *Proceedings of the Workshop on Hot Topics in Operating Systems* (Association for Computing Machinery 2021).

Ronald T Kneusel, *How AI Works: From Sorcery to Science* (No Starch Press 2024).

Joe Reis, Matt Housley. *Fundamentals of Data Engineering* (O'Reilly 2022).

Giovanni Sartor and Francesca Lagioia, 'The Impact of the General Data Protection Regulation (GDPR) on Artificial Intelligence' (European Parliamentary Research Service, 2020).

Joel Spolsky, 'The Law of Leaky Abstractions' in *Joel on Software: And on Diverse and Occasionally Related Matters That Will Prove of Interest to Software Developers, Designers, and Managers, and to Those Who, Whether by Good Fortune or Ill Luck, Work with Them in Some Capacity* (Apress 2004).

Unit 3. Cybersecurity Aspects of Artificial Intelligence

By the end of this unit, learners will be able to **distinguish** between three sources of cybersecurity risks from AI—the amplification of existing risks, new risks related to AI training and inputs, and new risks related to AI algorithms—and **combine** these sources for a more comprehensive examination of AI applications.

This unit discusses how cybersecurity concerns can emerge in contexts involving AI. By cybersecurity, we refer to the practices involved in protecting computer systems—in this case, AI systems and models—from deliberate interference by external actors. This interference can come from various sources and means of attack: disgruntled employees might want to leak a company's trade secrets, hackers might want to steal citizen data from a government body to commercialize it, hostile states might want to infiltrate government networks and steal intellectual property from companies, and so on. Over the past few decades, a sophisticated body of knowledge has been developed around cybersecurity. Our goal here is to briefly discuss how that body of knowledge relates to AI, both when it comes to known challenges that continue to exist in AI technologies and to novel issues that appear from what is unique about AI.

Cybersecurity, in AI as elsewhere, follows some core principles that guide the protection of systems against **unwanted interference**. The core principles of confidentiality, integrity, and availability, commonly referred to as the **CIA triad**, form the foundation for protecting information systems and data:

- The principle of **confidentiality** states that computer systems should prevent unauthorized individuals from gaining access to data. For example, one of the measures that **DigiToys** can take to promote confidentiality is limiting access to the data collected from its toys to the individuals who need to use that data in their work.
- The principle of **integrity** is that information should not be altered, either maliciously or accidentally, and that it must remain reliable for its intended use. Integrity is critical in contexts where decisions are made based on data analysis, including AI-driven systems. For instance, if the data used by **InnovaHospital** to make decisions about patient treatments is tampered with, medical resources might be misallocated, and individuals might be assigned to inadequate procedures.
- The principle of **availability** states that data and systems are accessible when needed, by authorized users. This is particularly important for AI systems that may operate in real-time or support critical functions, such as fraud detection or autonomous decision-making.

Other principles can be relevant in specific contexts. One such principle is **non-repudiation**, that is, the idea that a user should not be able to deny their involvement in an action or transaction. This principle is relevant, for example, in digital contracts or electronic payments, where it prevents individuals from falsely claiming that they did not sign a document or approve a transaction. The applicability of such principles is sometimes narrower than the CIA triad, but they might be no less important in their domains of use.

Together, these concepts illustrate a broader cybersecurity goal: the protection of data from unauthorized access, alteration, and disruption. While these principles are not new to data protection officers, their application within AI systems—where data flows, processing methods, and potential vulnerabilities are more complex—demands a nuanced understanding and an integration of both privacy and security frameworks.

To assist you with developing such an understanding, **Session 3.1** offers a refresher course of basic cybersecurity concepts and discusses how they become legal obligations under the AI Act. **Session 3.2** then revises general cybersecurity threats that can affect all forms of data processing, including the processing that happens in AI technologies. Finally, **Session 3.3** focuses on cybersecurity issues that are specific to AI.

Session 3.1. Core concepts and legal requirements for cybersecurity

> By the end of this session, learners will be able to **explain** what cybersecurity entails for data processing and **describe** some of the most common risks to it.

Threats to the cybersecurity principles discussed in the introduction to this Unit can take various forms. Each of those principles might be affected to a different extent by different practices aimed at different goals. To facilitate discussion of these issues, cybersecurity professionals have developed a shared vocabulary, as well as resources for the spread of knowledge.

Some of the better-known resources on cybersecurity are offered by the MITRE corporation to the general public, such as ATLAS (a knowledge base of adversaries and techniques used to attack digital systems) and D3FEND (a visualization of cybersecurity measures). In Europe, ENISA (the EU agency for cybersecurity) offers a broad set of tools that companies can use, such as best practices and self-assessment tools. It also publishes materials, such as guidelines, to reflect best practices in cybersecurity as well as risks that have become salient in an European context. Additionally, data protection authorities are also active in the cybersecurity domain, because, as we discuss below, security is an integral part of the protection of personal data.

In this session, we will discuss basic concepts that must be understood to make the best use of those resources. We will also cover the legal obligations that make cybersecurity a central requirement for legal compliance.

Approaches to cybersecurity

To pursue cybersecurity, organizations must take actions. Some of these measures are **reactive**, as they offer responses to security incidents after they occur. For example, if an organization discovers some of the personal data it stores has been stolen, it will often contact the affected individuals and offer them access to tools such as credit monitoring.

Reactive security involves activities like incident response, damage assessment, and remediation efforts to restore normal operations. While necessary, reactive measures are limited in their ability to prevent future attacks. For instance, a DPO might work with IT teams to address a data breach by securing affected systems and notifying regulators. Doing so can eliminate known issues that resulted in the data breach, but future attacks might still be possible from vulnerabilities that were not yet seen.

Other measures are **proactive**, as they seek to anticipate and prevent security issues before they arise. Proactive security includes regular vulnerability assessments, threat intelligence gathering, penetration testing, and implementing robust security policies. Proactive measures are particularly relevant in AI systems, where pre-emptive assessments of model security can help mitigate risks associated with adversarial attacks or data leakage. By identifying potential threats early in the life cycle of an AI system, organizations can implement safeguards to reduce the likelihood of a successful attack.

Most organizations will rely on both reactive and proactive measures to address their security challenges. A popular approach for determining the measures that are relevant in a context is that of drawing a **threat model**. Such a model offers a structured approach used to identify, evaluate, and prioritize potential security risks to a system, application, or data. A threat model typically outlines:

1. The attack surface, that is, the points where a system could be vulnerable to attack.
2. Potential threats or threat actors, such as hackers, criminal organizations, or nation-state attackers
3. The likelihood and impact of these threats.

Sometimes that information must be procured from outside an information, for example by tapping into the expertise of contractors. In other cases, it is already available within an organization but dispersed among many actors. It might be the case, for instance, that nobody has the full picture of how a particular AI system is designed and used. By

articulating all this knowledge, a threat model supplies a starting point for thinking about cybersecurity risks and how to respond to them.

To produce a plausible threat model, an organization must have a deep knowledge of both its technical tools and the context in which those tools are used. Based on that knowledge, an organization can anticipate potential threats and propose measures that will eliminate them, or at least mitigate the likelihood or severity of any attacks.

The attacker as the adversary

Cybersecurity, as mentioned above, refers to protection against deliberate efforts to affect a computer system. These deliberate efforts are made by an **attacker**, which is the term used to refer to any individual or entity attempting to exploit vulnerabilities to gain unauthorized access or cause harm. Thwarting the goals of attackers is necessary to ensure the cybersecurity principles discussed above. Following the cybersecurity principles, in turn, is valuable because it leads to other goals—such as the protection of the fundamental rights to privacy and data protection.

Identifying and classifying attackers can be complex, as their motivations, methods, and resources vary widely.

1. **Individual attackers** might include hackers driven by curiosity, personal grievances, or financial gain. They often use publicly available tools and exploit common vulnerabilities. For instance, a disgruntled former employee might use their retained access credentials to leak sensitive data as an act of retaliation.
2. **Criminal organizations** operate with more coordination and sophistication, often driven by profit motives. These groups may engage in activities like ransomware attacks, data theft, and fraud. In AI contexts, criminal organizations might target proprietary algorithms or large datasets used for training, aiming to steal valuable intellectual property or disrupt business operations.
3. **Nation-state attackers** are state-sponsored entities conducting cyber espionage, sabotage, or warfare. These attackers are typically well-resourced and highly skilled, targeting critical infrastructure, government systems, or large corporations for strategic gains. For example, an AI-based facial recognition system used for border control could become a target for nation-state attackers aiming to discredit the country deploying the system or ensure that their operatives can freely access that country.

Classifying attackers is not always straightforward, as their methods can overlap, and motivations may change over time. Moreover, the use of anonymization techniques, such as VPNs and the dark web, makes it challenging to trace the origin of attacks, complicating attribution efforts. Still, any organization’s threat models need to consider the kinds of resources that might be available to whoever wants to attack it.

Legal requirements for cybersecurity in the EU

Organizations might not always be aware of the cybersecurity risks they face, but they have a strong self-interest in avoiding those risks. Data breaches, intellectual property, and other security risks might have an unbearable financial cost to businesses. Even for public sector entities and non-profits, cybersecurity issues might erode the organization's legitimacy or disrupt its ability to do its job, as seen in the constant ransomware attacks that have become common in recent years. If self-interest is not enough, many organizations are also subject to legal obligations to pursue cybersecurity.

For the purposes of this training module on AI and data protection, our focus will reside on cybersecurity requirements in the GDPR and the AI Act. Article 32 GDPR requires any data controllers to adopt technical and organizational measures to ensure that their data processing has a level of security compatible with the risk associated with it. For high-risk AI systems, Article 15 AI Act obliges providers to ensure the system has a level of cybersecurity appropriate to its purpose. In both cases, the obligations apply at the moment an AI system is designed and also when it is effectively used to process personal data.[1] Those legal requirements are discussed throughout the module.

Cybersecurity requirements in the GDPR and the AI Act coexist with other legal instruments in this domain. Sector-specific legal instruments, such as the Medical Devices Regulation, can feature specific standards for particular applications of AI technologies. Additionally, the EU has adopted various legal instruments on cybersecurity, which establish additional rules. Under the NIS2 directive (Directive (EU) 2022/2055), for example, Member States are obliged to establish legal requirements for cybersecurity in systems used for certain applications. More generally, the recently adopted Cyber Resilience Act establishes essential security requirements that must be observed for placing products with digital components in the EU market, including high-risk AI systems. The training module will not go into the details of those requirements, but organizations will need to consider them when deciding how to fulfil their cybersecurity obligations under the GDPR.

[1] See Unit 12 of this training module.

Session 3.2. General threats to cybersecurity

By the end of this session, learners will be able to **give examples** *of* cybersecurity threats and their impact on the protection of personal data. They will also be able to **exemplify** best practices to reduce the risk from those threats.

This section provides an overview of the cybersecurity risks that affect computer systems in general. The concepts and practices discussed here go beyond AI, as they might affect all kinds of software. Still, AI systems and models remain vulnerable to them. Organizations developing or deploying AI technologies cannot ignore these threats just because they are not AI-specific. As such, it will be important to review general issues of cybersecurity before moving on, in the next session, to the unique AI-related risks to cybersecurity.

Cybersecurity, as defined in this unit, is concerned with deliberate practices. It might be threatened by **attacks**, in which an individual, group, or organization tries to breach the security of the information system, network, or digital device in question. In the previous unit, we have seen that attackers might have a variety of profiles, resources, and goals in their attacks. In particular, they can target any of the aspects of the CIA triad—confidentiality, integrity, and availability.

A **security vulnerability** is a weakness or flaw in a system, software, or process that can be exploited by an attacker to gain unauthorized access, cause disruptions, or compromise data. Vulnerabilities can arise from coding errors, misconfigurations, outdated software, or even insecure design choices. For instance, a web application vulnerability like SQL injection could allow an attacker to manipulate a database and access confidential information, such as user credentials or payment data. In the context of AI systems, vulnerabilities may include poorly secured training data, biased algorithms, or exposure of sensitive data through model inversion attacks.

A **zero-day vulnerability** refers to a security flaw that is unknown to the software vendor and, therefore, unpatched. Attackers who discover a zero-day exploit have a significant advantage, as there is no immediate fix available to prevent exploitation. For example, a zero-day attack on a popular cloud service provider could enable attackers to infiltrate customer data before the vulnerability is detected and patched.

A **security incident** is any event that compromises the confidentiality, integrity, or availability of information or systems. Incidents can range from minor breaches, such as unauthorized access to a single user’s email account, to major data breaches affecting millions of individuals. The impact of a security incident can be severe, often requiring

incident response measures, reports to regulators,[2] and actions to prevent recurrence. For AI-driven systems, a security incident might involve unauthorized manipulation of model behaviour, such as an adversarial attack that misleads an image recognition system into misclassifying objects. We will now look at some of the approaches attackers use for creating security incidents.

Types of attacks

One can distinguish between two main types of attacks. An **active attack** happens when an attacker directly interferes with the computer system in question. For example, they might manipulate the data that is used for training an AI system or use carefully designed prompts to "jailbreak" a large language model, that is, to extract information and parameters from a model. **Passive attacks**, instead, do not engage directly with the system but monitor its operation. For example, an attacker might monitor all the requests that are sent to a given AI system in order to better understand how that system is used. Attackers often rely on both approaches, which can be applied in various forms.

The most prevalent attack methods often exploit human behaviour, software vulnerabilities, and weaknesses in data transmission processes. As innovative technologies emerge, they might be vulnerable to new ways to carry out attacks. At the same time, cybersecurity practitioners might develop methods that eliminate or reduce the risk from certain attacks. Because of this arms race, it is difficult to keep track of the diversity of attacks used by malicious actors. Resources such as MITRE's ATLAS knowledge base offer a shared repository of knowledge on the current state of the art. Based on that knowledge, one can group attacks into some classes that remain relatively stable over time, even if the details of their implementation vary wildly.

Social engineering

Social engineering is a technique that exploits human psychology rather than technical vulnerabilities. Attackers manipulate individuals into divulging confidential information, such as login credentials or sensitive personal data. Common forms of social engineering include phishing, where attackers send fraudulent emails that appear legitimate, tricking recipients into clicking malicious links or providing sensitive information.

A phishing email may masquerade as a message from a bank, asking the user to reset their password through a provided link. Once the user enters their credentials on a fake website, the attacker can use those credentials to gain access to their target system. In the context of AI, social engineering attacks might target employees with access to sensitive training data or AI model configurations, compromising the system from within.

[2] As required, for example, by the GDPR.

Exploiting software vulnerabilities

Attackers can also proceed by exploiting known vulnerabilities. An **exploit** is a piece of software, script, or code designed to take advantage of a security vulnerability in a system or application. When attackers discover such a weakness, they can use an exploit to gain unauthorized access or execute malicious commands. For example, a buffer overflow exploit targets a vulnerability where a program fails to properly check the length of input data, allowing an attacker to overwrite memory and execute arbitrary code. In AI applications, exploits might focus on software libraries used for machine learning, compromising the integrity of the model, or extracting sensitive information from the system.

Man-in-the-middle attacks

A **man-in-the-middle (MITM)** *attack* occurs when an attacker secretly intercepts and alters the communication between two parties who believe they are directly communicating with each other. In this scenario, the attacker positions themselves between the sender and receiver, allowing them to eavesdrop on, modify, or inject malicious content into the data exchange.

In an unencrypted Wi-Fi network, for example, an attacker can intercept data sent between a user's device and a web server, capturing sensitive information like login credentials or financial details. In AI systems, MITM attacks can disrupt the transmission of data used for model training or inference, potentially introducing false data inputs that lead to incorrect outputs or compromised decision-making.

Putting it all together

While these forms of attacks are distinct, they are often used in combination by attackers to increase their chances of success. For example, an attacker might use social engineering to gain initial access, exploit a software vulnerability to escalate privileges, and then carry out a man-in-the-middle attack to intercept and manipulate data. In AI environments, the complexity of interconnected systems and the reliance on large datasets can amplify these risks, as attackers may target weak points in the data pipeline or leverage adversarial inputs to compromise model integrity. This is why the AI-specific threats discussed in Session 3.3 of this training model cannot be separated from the more established attack vectors seen here.

Types of security controls

Security controls are measures designed to protect information systems from threats and reduce risks. These controls can be classified into distinct categories based on their primary function:

1. **Preventive Controls** are aimed at stopping security incidents before they occur. This includes measures like firewalls, encryption, access controls, and multi-

factor authentication. For example, encrypting data at rest and in transit ensures that even if an attacker gains access, the data remains unreadable without the decryption keys.

2. **Deterrent Controls** are intended to discourage potential attackers from attempting to exploit a system. These controls might involve visible security measures, such as warning banners, surveillance cameras, or legal disclaimers about monitoring and prosecution. In the context of AI, deterrence might include transparent declarations of robust model validation processes, signalling to potential attackers that their efforts are likely to be detected.
3. **Detection Controls** focus on identifying security incidents as they happen. Examples include intrusion detection systems (IDS), anomaly detection algorithms, and security information and event management (SIEM) tools. For AI systems, detection controls might involve monitoring inputs for unusual patterns or adversarial attacks designed to manipulate model outputs.
4. **Deflection Controls** aim to divert attacks away from critical systems, often by misleading attackers. This can involve the use of honeypots—decoy systems designed to attract and trap attackers, giving security teams time to respond. For instance, setting up a fake server that mimics a valuable database can lure attackers away from the real system.
5. **Mitigation Controls** seek to limit the damage caused by a security incident. These include measures like data backups, network segmentation, and incident response plans. In AI systems, mitigation might involve reverting to a safe fallback model if anomalous behaviour is detected, reducing the impact of compromised algorithms.
6. **Recovery Controls** help organizations return to normal operations after a security incident. These measures include data restoration, system reboots, and process reviews to prevent future occurrences. Effective recovery controls are essential for minimizing downtime and ensuring business continuity, especially in AI applications that support critical functions like financial transactions or healthcare diagnostics.

Various kinds of controls are often used together. The concept of **security in depth** advocates for a multi-layered approach to cybersecurity, where multiple, overlapping controls work together to protect systems and data. This strategy recognizes that no single control is foolproof; instead, various measures complement each other to create a more robust defence. For example, an organization might use a combination of firewalls, intrusion detection systems, data encryption, and user access controls to secure its infrastructure. This approach is often illustrated with the cheese layers model.

In the cheese layers model, each layer of defines is depicted as a slice of cheese with holes (representing vulnerabilities). While a single layer may have weaknesses, the holes rarely align perfectly across multiple layers. Thus, if one control fails, another layer can still block the attack. For instance, even if an attacker bypasses the firewall, they may still be detected by the intrusion detection system. This layered defence strategy is crucial for AI systems, where the many potential points of failure, such as model vulnerabilities and data privacy risks, require diverse and adaptive controls.

The thin line between security practices and cyberattacks

Sometimes, it can be difficult to distinguish between attack practices and the cybersecurity practices used to address them. One example comes from the recent growth in fuzzing methods. **Fuzzing** is a technique used to find software vulnerabilities by providing random, unexpected, or invalid data inputs to a program and observing its behaviour. The goal of fuzzing is to identify weaknesses that can be exploited by attackers, such as crashes, memory leaks, or unexpected behaviour that indicates poor input handling.

For example, a fuzzing tool might send a series of malformed inputs to a web application in an attempt to trigger a vulnerability like a buffer overflow or input validation error. In AI systems, fuzzing can be used to test the robustness of machine learning models, identifying scenarios where the model fails or produces unreliable results due to unanticipated input patterns.

Fuzzing is largely mentioned as a cybersecurity tool, which organizations use to anticipate attacks that might be used against them. In this context, AI technologies can be used to boost cybersecurity, by allowing cybersecurity experts to create and test a larger number of scenarios. However, the same techniques to detect vulnerabilities might be used by an attacker who wants to figure out how to actively attack a system. If that happens, the use of AI systems increases the capabilities of attackers. This means that AI technologies do not end the arms race between attackers and defenders but continue to feed it.

Session 3.3. AI-specific risks to cybersecurity

By the end of this session, learners will be able to **indicate** how the use of AI creates unique risks from a cybersecurity perspective.

AI systems and models are complex objects, as we have seen in Unit 2. This means that an attacker has a wealth of points they can probe for potential vulnerabilities. An attack might target the data used to create an AI system, its training process, the infrastructure used to support its execution, or its context of use. At each juncture,

various methods can be used to identify and exploit vulnerabilities. In this session, we focus on attacks that are specifically tailored for AI systems and models.

Given the current predominance of machine learning models, this session will mostly deal with attacks directed at machine learning technologies. Our goal here is not to discuss the intricacies of those attacks, as many of them rely on technical elements that require some expertise. Learners who want a bit more of technical detail would do well to consult other materials, such as the **Elements of Secure AI Systems** training module for ICT professionals. Instead, we will focus on present the general features of those attacks, so that data protection professionals can collaborate with technical experts in raising awareness about them and designing organizational responses.

One thing that must be kept in mind, however, is that cybersecurity in AI is a relatively novel domain. As such, attackers are often in a more advantageous position in comparison with defenders. They only need one successful exploit of a vulnerability, whereas a defender needs to clear all risk vectors. However, because the AI techniques themselves are novel, sometimes there are no known ways to fully eliminate the risk. Therefore, organizations will sometimes be forced to evaluate whether existing measures for mitigation can reduce risk to a legally acceptable level. Otherwise, they might be forced to abandon the use of AI for that specific purpose.

Attacks on the AI training process

AI models, particularly machine learning systems, can be subject to cybersecurity threats during their training stage. Those threats might impact various desirable properties of AI systems. Consider the CIA triad:

1. **Confidentiality** is relevant at the training stage, as organizations might want to preserve their expertise codified in the model and training practices, and they remain subject to data protection requirements that require them to control access to any personal data used in training.[3]
2. When it comes to **integrity**, AI systems and models rely heavily on large datasets to learn patterns and generate their outputs. This is often summarized in the maxim “garbage in, garbage out”: if one starts from bad training data, the ensuing model is likely to be inaccurate or even misleading in important ways. As a result, the integrity of training data becomes crucial for model performance and reliability.
3. **Availability** issues are a bit less salient at the training stage, but they might still occur, for example, when a model continues to learn after it is deployed.

[3] See Unit 6 of this training module.

As discussed throughout the unit, those goals can be affected in many ways. We should now consider attack vectors that are specific to AI.

One major risk in the training phase is **data poisoning**, where attackers intentionally manipulate the training data to influence the behaviour of the model. For example, a hacker might introduce mislabelled exams into **InnovaHospital**'s databases. If trained on those mislabelled exams, an automated diagnosis algorithm might clear patients that are in fact sick or provide false positives to healthy patients. Data poisoning is particularly concerning in scenarios where the training data comes from external or crowdsourced sources, as these datasets are more susceptible to tampering.

A variant of data poisoning is the so-called **backdoor** *attack*, in which the model training is sabotaged to ensure that a model produces an incorrect output when it identifies a certain element in the input. Consider a scenario where **UNw** decides to adopt an AI system for automatically grading undergraduate exams. A malicious student, knowing about this, hacks into the system's training data and inserts data that falsely labels any exams taken by them or their friends as receiving the highest grades. This would allow these students to perform well regardless of their actual effort.

Attackers can also tamper with an AI system through **environmental attacks**. In this kind of attack, the system itself is not altered, but the attacker directs their attention to the environment in which the system will operate. For example, a malicious competitor of **DigiToys** might compromise software libraries that are used by this company, with a view to making their AI systems not working or introducing backdoors for exfiltrating information. As we discussed in Session 2.3 of this training module, the training of AI models depends on a complex environment. Hence, attackers have many opportunities to exploit vulnerabilities in different pieces of the infrastructure supporting an AI system.

Attacks on deployed AI systems

Once an AI model has been trained and deployed, it remains vulnerable to a diverse set of attacks that target its predictions and outputs. As more AI models are used in a variety of real-world applications, attackers can identify new vulnerabilities they can exploit. Those vulnerabilities can take many forms, many of which rely on interactions with the AI system.

One common attack against deployed models is the **adversarial attack**. In this approach, attackers carefully craft input data designed to deceive the AI model. For example, an adversarial image might appear normal to a human observer but contains subtle perturbations that cause a computer vision model to misclassify it. This type of attack could be used to trick facial recognition systems into misidentifying individuals or to manipulate AI models used in autonomous vehicles, potentially leading to dangerous consequences.

Deployed AI systems might also be vulnerable to **model extraction** attacks. In this kind of attack, the malicious party attempts to replicate a deployed AI model by querying it extensively and gathering information about its outputs. Through repeated interactions, the attacker can approximate the decision-making process of the original model. For example, they might obtain information about which safeguards have been implemented in the model and which values have been given to certain key parameters.

Model extraction attacks are particularly problematic for proprietary AI models that represent significant investments in research and development. The stolen model can then be used by competitors or malicious actors, undermining the original creator's competitive advantage. More generally, however, a model extraction model can be a starting point for further exploitation. An attacker might simply want to duplicate the extracted model for their own purposes, but they might be interested in carrying out further attacks. In the latter case, access to an extracted model will allow them to identify other vulnerabilities that can be used for follow-up attacks.

To conclude this necessarily incomplete overview of attacks against deployed AI systems, we must talk about a third kind of attack—**model inversion**. Just like model extraction attacks, model inversion operates by repetition. The attacker makes various queries to the AI system and uses the system's outputs to extract information from it. In this case, however, the goal is not to extract the model itself, but data used during its training process. For instance, if an AI model is trained on a medical dataset, model inversion techniques could potentially reveal private details about individual patients. This means model inversion attacks can directly affect the level of protection afforded to personal data used for training AI.

Challenges in addressing AI-specific cybersecurity risks

The fast pace of evolution of AI technologies creates a tough cybersecurity challenge. Innovative technologies give origin to evolving threats, at the same time there are few measures that have been proven to be effective in mitigating or eliminating risks. Deploying those techniques sometimes requires advanced expertise of a different kind than the one used for developing and deploying AI systems. Furthermore, AI technologies themselves can be leveraged for detecting and exploiting vulnerabilities. Even so, there are various actions that data controllers can take when it comes to AI-related risks.

One of the fundamental challenges in AI cybersecurity is the evolving nature of the threats and the limited availability of proven mitigation measures. Some AI systems continuously learn and adapt, which introduces new attack surfaces. Furthermore, the complexity and opacity of many AI models make it difficult to understand their vulnerabilities fully. This lack of transparency, often referred to as the "black box"

problem,[4] complicates efforts to identify potential weaknesses and implement robust defences.

Mitigating these attacks is complex, as many standard defences are not fully effective against the attacks outlined in this session. Techniques like input validation, adversarial training (where the model is exposed to adversarial examples during training), and rate-limiting of model queries can help, but they are often insufficient. Adversarial attacks, in particular, highlight the fragility of AI models, as even small, imperceptible changes to input data can lead to incorrect outputs. Additionally, the lack of effective countermeasures against model extraction creates a significant risk for public-facing applications of AI, especially those that require sensitive data to work.

In the absence of AI-specific solutions for AI-specific issues, data controllers need to rely on established cybersecurity approaches. Defence in depth approaches—which rely on multiple, overlapping measures to diminish risk and safeguards to deal with harm—can compensate the shortcomings of individual techniques. This means organizations need to look at measures directed at different components of an AI system or model, taking effect throughout its entire life cycle.[5] Even so, certain risks may remain unaddressed due to the novelty and complexity of AI-specific attacks. Therefore, defence in depth is not a silver bullet for AI cybersecurity.

In some cases, the risks associated with deploying an AI system may outweigh the potential benefits. This is particularly likely to be the case when the system handles sensitive data or is used in critical decision-making processes. For instance, using AI in healthcare diagnostics or for criminal investigations may introduce unacceptable risks if the models are vulnerable to adversarial attacks that could lead to incorrect or harmful outcomes. In such situations, organizations might consider alternative approaches, such as relying on simpler, rule-based systems or employing a hybrid approach where AI decisions are supplemented by human oversight.

Conclusion to Unit 3

The unique cybersecurity risks posed by AI technologies require a careful balance between fostering innovation and ensuring robust risk management. As the field of AI security is still in its initial stages, there are limited standardized solutions for many of the emerging threats. This creates a challenging environment for organization, who must navigate the complexities of AI risk to comply with their data protection obligations. The safe deployment of AI technologies thus requires collaborative efforts between data protection experts, AI developers, and cybersecurity professionals.

[4] See Session 4.3 of this training module.
[5] For a closer look at the idea of AI life cycle, see Part II of this training module.

Based on the previous discussions, data protection professionals would do well to keep some points in mind during their assessments:

- Having **clear threat models** for a given AI application or model can help in the diagnosis of potential risks.
- AI systems remain vulnerable to many attacks that affect software systems in general. Therefore, AI cybersecurity needs to attend both to the AI models and to the **non-AI components** that allow their use.
- AI technologies can be used both by organizations in identifying and responding to cybersecurity vulnerabilities and by attackers in exploiting those vulnerabilities.
- Currently, the novelty of AI technologies favours attackers rather than defenders. There are **no known measures to respond to certain risk vectors**.
- Organizations need therefore to **consider whether existing measures and safeguards can reduce risks** to an acceptable level.
- If risk can be reduced, a **defence in depth approach** might help overcome the limitations of individual AI cybersecurity techniques.
- Otherwise, an organization might need to consider whether it can lawfully deploy AI at all if it cannot ensure a minimum level of cybersecurity.

Ultimately, the integration of AI into data processing and decision-making processes requires a shift in the traditional approach to cybersecurity. Data protection professionals must adopt a proactive stance, focusing not only on compliance but also on the broader implications of AI risks. Effective data protection in the age of AI will therefore require close attention to a technical landscape that is both extraordinarily complex and fast-moving. This, in itself, is not different from usual data protection practices. But the specific technical arrangements of AI can make much difference for whether and how problems can be addressed.

Prompt for reflection

Based on **DigiToys**' focus on reputation and compliance, what proactive cybersecurity measures could they prioritize to mitigate AI-specific risks in their products?

References

Ross Anderson, *Security Engineering: A Guide to Building Dependable Distributed Systems* (3rd edn, Wiley 2020).

Federica Casarosa, 'Cybersecurity of Internet of Things in the Health Sector: Understanding the Applicable Legal Framework' (2024) 53 Computer Law & Security Review 105982.

Markus Christensen and others (eds.), *The Ethics of Cybersecurity* (Springer 2020).

Henrik Junklewitz and others, *Cybersecurity of Artificial Intelligence in the AI Act: Guiding Principles to Address the Cybersecurity Requirement for High Risk AI Systems*. (Publications Office of the European Union 2023).

Andrei Kucharavy and others (eds), *Large Language Models in Cybersecurity: Threats, Exposure and Mitigation* (Springer 2024).

Taner Kuru, 'Lawfulness of the mass processing of publicly accessible online data to train large language models' (2024) International Data Privacy Law.

MITRE Atlas.

MITRE D3FEND Matrix.

Kaspar Rosager Ludvigsen, 'The Role of Cybersecurity in Medical Devices Regulation: Future Considerations and Solutions' (2023) 5 Law, Technology and Humans 59.

Unit 4. The Safe Use of Artificial Intelligence

By the end of this unit, learners will be able to **analyse** risks associated with the development and use of AI, **distinguishing** between cybersecurity risks, risks related to failures in functionality, risks related to the effects produced by an AI system, and risks from opacity.

Data protection law aims to protect the fundamental rights and freedoms of natural persons from the risks that might come from data processing.[1] Some of these risks, as we have seen in the previous unit, emerge from deliberate attempts to interfere with computer systems or extract information from them. Others, however, emerge from the operation of the systems themselves. Computer systems can fail in their operation, or they might produce undesirable side effects even if running correctly. For example, scholars and activists have recently pointed out various environmental hazards coming from the growing use of AI. In this unit, we will discuss some aspects of AI technologies that can affect their safe development and use from a data protection perspective.

The concept of safety is a complement to the concept of security we discussed in the previous unit. They both relate to the prevention of harms coming from a computer system. However, they cover distinct kinds of harm. Whereas security is concerned with preventing malicious interferences with a computer system, **safety refers to the prevention of harms that do not involve an attacker** (Herrmann and Pridöhl 2020). Those harms might be the result of natural events, such as a storm that disrupts the operation of a system that maintains a critical piece of infrastructure. Or they might be the result of the system's operation: for example, an AI chatbot designed to defraud users will harm those users if it functions as expected. This means that an AI system must be both safe and secure to comply with legal requirements.

Safety is a complex phenomenon. It relates to social, psychological, and institutional factors, among others (Levenson 2012). As a legal obligation, it can find a wide variety of sources. One of those is data protection law. Under Article 25 GDPR, data controllers are obliged to take technical and organizational measures that consider the risks that processing can create for the rights and freedoms of natural persons.[2] This means that lawful processing requires attention to not putting those rights and freedoms at stake.

In this unit, we will examine three sources of risk to safety that are particularly relevant in the context of AI. **Session 4.1** discusses numerous factors that can make an AI system's actual operation diverge from the promises used to sell it. **Session 4.2** then

[1] Article 1(2) GDPR.
[2] Note that the provision does not speak of "fundamental rights." As such, it covers the broader range of legally recognized interests that an individual might have.

illustrates how some risks can emerge even if an AI system operates as expected. Finally, **Session 4.3** discusses how factors such as the technical complexity of AI systems, their scale of operation, and intellectual property rights can be obstacles to the evaluation of AI systems.

Session 4.1. The promise of functionality and its limits

> By the end of this session, learners will be able to **illustrate** reasons that might lead an AI system to not operate as expected, such as defects in software design, biased algorithms, or inadequate organizational processes.

Whenever somebody uses or creates an AI system or model, they usually intend it to have one or more functionalities. That is, there is an expectation that the AI technologies are used to do *something*. For example, a chatbot is expected to be able to interact with humans in conversations, while a facial recognition system is expected to recognize faces. Yet sometimes these functionalities are not actually present in the finalized system or model. Or, if they are, the AI technology performs worse than expected. In this session, we will discuss how that can happen and why that matters for data protection.

Over the past decade, the speedy developments of AI technologies create expectations that AI can solve any problems. Even if no technology available today can solve a given problem, this might not be the case in a few years. For example, the object recognition capabilities that are available in a moderately-priced smartphone nowadays were beyond the reach of computer science only a decade ago. As such, the adoption of AI technologies is driven not just by our knowledge of what we know AI can do for sure today, but by the **promise** that certain technologies show of solving future problems (Hirsch-Kreinsen 2024).

These promises do not always materialize in practice.[3] Back in the 1950s, computer scientists had expected to solve most of major technical problems behind AI in a summer. Technologists and entrepreneurs keep promising that we will have super-intelligent AI systems, self-driving cars, and other technologies, and they keep revising their estimates of when those technologies will actually be available. Even more modest promises often fail to materialize: IT projects are notorious for taking much more time and effort than original forecasts (see, e.g., McConnell 2006). As such, data protection professionals would do well not to take the promises of software development at face value.

[3] For a study of overpromising in technology, see Gaillard et al. (2023).

Analysing how things can go wrong

One challenge that we face when analysing safety issues is that **many things can go wrong**. To approach this problem, one can follow a similar approach to the one adopted in cybersecurity: relying on **shared knowledge** bases that catalogue potential sources of unsafety in the development and operation of an AI system. Some initial steps towards this have been taken, as organizations such as the OECD have created databases that monitor safety incidents related to artificial intelligence. By sharing reports of those incidents, individuals and organizations can understand and draw lessons from what has gone wrong.

To systematize the lessons learned from AI safety incidents, one might expand on them and offer some theoretical constructions. A potentially fruitful approach for this has been proposed by Deborah Raji and her co-authors. In a 2022 conference paper, these authors identify what they call the **fallacy of functionality**: the mere fact that an AI system exists is not enough for us to believe that it does what it promises to do. This is because AI systems can fail in many ways.

Beyond this concise formulation of the fallacy, Raji and her co-authors offer a taxonomy of failure modes of AI. That is, they classify several ways in which an AI system might fail to deliver the promised functionalities. In the following paragraphs, we will look more closely at the categories proposed by those authors.

The first type of failure mode they cover is that of **impossible tasks** (Raji et al. 2022, p. 962). Sometimes, an AI system cannot do what it is expected to do because that goal cannot be achieved at all. A task might be impossible at a conceptual level, for example if it tries to make predictions with no scientific basis, as is the case of various AI systems attempting to infer traits of personality, behaviour, or social status from physical traits (Stark and Hutson 2022). Other tasks are possible in theory but cannot be achieved in practice. Raji et al. (2022) give numerous examples of trying to build AI systems when the data available is biased or does not capture key features for the problem at hand.

Other failure modes stem from **engineering failures** (Raji et al. 2022, p. 963–964). AI systems and models are developed by humans, either acting alone or as part of larger groups and organizations. The individuals and groups working in an AI system are fallible, and this can affect the functionality of an AI system. They might fail to implement certain features correctly, to detect errors in the system, to include safeguards to individual rights and so on. For example, a programmer might use an outdated version of a software library when developing an AI system, one that gives wrong results for one of each one hundred analyses done by the system. Those errors in programming might produce harms to individuals, for example, if they lead to an individual being assigned the wrong treatment by a medical AI system.

The third category in the failure taxonomy is that of **deployment failures** (Raji et al. 2022, p. 964). These failures refer to various things that can go wrong when one puts an AI system to use:

1. A system might lack **robustness**; that is, its outputs might be disturbed if the conditions in which it is used change just a little. For example, a robust AI system for evaluating student performance should not change its predictions radically if a student's grade in a specific exam is revised a few decimal points up or down.
2. A system might suffer from **adversarial attacks**, as discussed in the previous unit, which are meant to interfere with its operation.
3. A system might fail to account for **unexpected interactions**. For example, a medical AI system used to diagnose heart diseases by looking at chest images might struggle if it is exposed to a patient with *situs inversus*.[4]

Those problems might affect even an AI system that has been well-designed to achieve a feasible task.

Finally, Raji et al. (2022, p. 964-965) discuss **failures of communication**. One example they give are situations in which a vendor overstates or even falsifies the capabilities of a technology they are selling. For example, a provider of a chatbot might claim that their systems can reason, while the system is actually just dealing with statistical correlations. The other example of communication failure they mention is that of misrepresented capabilities, which can happen if a provider sells a product even if they know it cannot be reliably used for a certain application. In those cases, the problem is not so much on the technical object as it is on the communication between those offering the AI-based tools and those buying their promises.

Dealing with the fallacy of functionality

From an organizational perspective, those who use AI technologies should benefit from looking closely at the failure modes mentioned above. Otherwise, they might find themselves buying (or even developing) tools that do not do what they promise, and so become expensive failures. However, addressing those fallacies is also a legal obligation when the AI systems are covered by data protection law.

For organizations deploying AI systems developed by others, the obligation follows from the fact that they will effectively be the **data controllers** for those systems.[5] As such, they must discharge various obligations towards the persons whose personal data is processed. If a system fails to operate as expected, the deploying organization will need to respond for any harms that failure might have caused. This means it will need to have

[4] A condition in which an individual's visceral organs develop in a mirror image of what is usually the case for humans, putting their heart in the right-side of the chest.
[5] See Unit 6 of this training module.

a clear view of what its AI systems can (and cannot do), otherwise the use of AI might expose it to undesirable liability.

For organizations developing or commercializing AI models or systems, the data protection obligations do not refer directly to the harms stemming from the use of those technologies.[6] Still, obligations can stem from other sources. An organization that develops an AI system is likely to be the data controller for any processing that takes place during the training process, and as such it will be responsible for safety failures. It might also have obligations of fair representation of its products. For example, the AI Act mandates various kinds of disclosure across the supply chain for providers of high-risk AI systems[7] and of general-purpose AI models.[8] A failure to critically engage with the fallacy of functionality might therefore lead to legal problems down the road.

Session 4.2. Adverse effects of AI applications

By the end of this session, learners will be able to **examine** how AI systems can harm the rights and interests of individuals and groups even if a system works as advertised.

In this session, we will discuss how AI systems and models might be unsafe even if they deliver all the promised functionalities. This is because many AI-based technologies are used in contexts in which they affect the physical and virtual environments where social life takes place. For example, online platforms often rely on content moderation algorithms, while governments might use AI systems to allocate benefits or detect fraud. The effect of AI systems in those use cases is not solely a function of their technical properties. Instead, it depends on the role those systems are expected to play and how they are operated within a given context.

Because the kind of harm we discuss here is sensitive to the contexts in which AI systems are used, it is not possible to cover all relevant cases. Instead, we will use the three hypothetical cases of Session 1.3 to illustrate how those harms might emerge in practice.

These examples highlight that the impact of AI systems extends beyond their functional performance. The broader context in which they are embedded often determines whether they contribute positively or negatively to society, particularly when it comes to protecting the rights and interests of individuals and groups. By looking at potential

[6] See, however, Session 6.1 of this training module on the possibility of joint controllership.
[7] See Article 13(2) AI Act.
[8] See Article 53 AI Act.

harms in each of the three case studies, and discussing their legal implications, our goal is to have some examples of factors that one can analyse in their own organization.

AI-based harms at the University of Nowhere

The **UNw** university is considering integrating AI technologies to alleviate the workload on its overburdened staff, particularly in administrative processes and student services. However, even well-functioning AI systems can lead to unintended harms that affect the rights and interests of students and staff. Let us now consider a few of the potentially harmful applications.

The use of AI-based systems for grading assignments and exams could introduce biases that disproportionately affect certain student groups. Automated assessment tools might systematically disadvantage students who come from non-traditional educational backgrounds, use unconventional writing styles, or whose first language is not the one used for instruction. For example, there have been various reports that AI-powered plagiarism detectors used in English-language institutions are more likely to wrongfully flag a student as a plagiarism if English is not their first language.

Another area of concern is the use of predictive analytics to identify students who may be at risk of dropping out. AI models could analyse student data, such as attendance records, grades, and engagement metrics, to flag individuals for intervention. While this may appear beneficial, it can also lead to privacy invasions and undue stress for students who are unfairly labelled as "at-risk" due to factors that the model misinterprets or oversimplifies. For example, students who work part-time jobs or have caregiving responsibilities might show lower engagement metrics but do not necessarily require or want additional intervention. This type of profiling can harm the students' sense of autonomy and increase stigmatization.

Processes such as those can harm the students affected by them. A dedicated student might be unfairly accused of plagiarism and have to spend time they would dedicate to studies in defending themselves against the charges. A student who is balancing their studies with full-time work to sustain their family might be required to follow remedial classes they have neither the need nor the time for. Such outcomes are not only unfair to the students but can lead to legal liabilities for the university.

From a data protection perspective, the following units of this course will help you identify various potential sources of non-compliance. Some of those are quite technical, but others can be identified from the incompatibility of these errors and biases with some GDPR principles. For example, the biases above run afoul of the principle of accuracy,[9] as they lead the university to store assessments about individuals. Any applications which make decisions about students without human involvement are also

[9] Article 5 GDPR. On data protection principles, see Session 6.3 of this training module.

likely to trigger the rules on automated decision-making from Article 22 GDPR. Furthermore, many of the applications outlined above are covered by the list of high-risk AI systems in Annex III AI Act, triggering additional obligations. As such, the potential impact of AI in students is something that must be considered beyond the technical rigour in design.

The risks of smart toys at DigiToys

DigiToys aims to use AI to create interactive, educative experiences for young children. Even if the AI embedded in the toys functions exactly as intended—engaging children with personalized learning prompts or responding accurately to their voice commands—there are still significant risks related to privacy and child development. Those risks are particularly relevant from a data protection perspective, as Recital 38 GDPR clarifies that the vulnerabilities of children warrant special protection when their data is processed.

As a recital, this stipulation is not legally binding. However, it points out how one should interpret the applicable legal provisions of the GDPR—not just the specific requirements for children's consent in Article 8, but *any* provision when a child's data is processed. In addition, other provisions of EU law also require special attention and protection to the rights and interests of a child, and the GDPR has to be interpreted in a way that is compatible with those.[10] Special attention to children is not just a desirable feature of the law, but a legal requirement, even if data controllers retain considerable flexibility on how to deal with that requirement.

Some of the challenges to children's rights might be directly connected with their right to data protection. For instance, if the toys track children's interactions to personalize the learning experience, they may inadvertently gather information about the child's behaviour, preferences, or even their emotional state. Part of that data might fall into the special categories of personal data defined in Article 9 GDPR, triggering additional requirements for processing.

Even if the gathered data is not deemed sensitive in a narrow legal sense, it can still pose considerable risks. Data collected from children might be subsequently processed for reasons that are not in their best interest, such as the creation of profiles from an early age. Those profiles might fail to consider how the interests, preferences, and even central aspects of a child's personality can change radically over time. For example, they might take into account mistakes that people make when they are young, even after those individuals have matured. This can adversely affect those individuals in adult life, and it might create obstacles from the exercise of their right to be forgotten.

[10] See, in particular, Article 24 of the EU Charter of Fundamental Rights.

The long-term implications of processing are not limited to **DigiToys**. The information gathered by that company might be shared with partner companies and organizations, requested by government authorities if national or EU law allows so, or even commercialized to data brokers. The spread of children's data might happen even if the company is deliberately averse to that: for example, if **DigiToys** goes bankrupt, its assets might be bought by companies that are less invested in child's rights.

Moreover, the use of AI in toys can alter the way children engage with the digital world, potentially affecting their cognitive and social development. Even if the toy is designed to be educational, there is a risk that children may become overly reliant on interactive digital stimuli, reducing opportunities for free play and human interaction. This can have long-term consequences on their ability to develop essential social skills, even if the toys are technically operating as intended.

Some challenges to automated medicine at InnovaHospital

InnovaHospital is known for its commitment to patient confidentiality and its embrace of innovative technologies. The integration of AI tools into clinical decision-making, such as diagnostic support systems or patient monitoring, may seem like a natural progression. However, even when these systems operate correctly, they can still produce harmful effects.

For example, an AI-based diagnostic tool might prioritize efficiency and speed, recommending standardized treatment protocols based on data-driven insights. While this may streamline care, it can also lead to a "one-size-fits-all" approach, overlooking individual patient needs or ignoring subtle symptoms that do not fit typical patterns. This could harm patients with rare conditions or those from underrepresented demographic groups whose medical data is not adequately represented in the training datasets.

Additionally, the use of AI in triaging patients could inadvertently exacerbate healthcare inequalities. An AI system designed to allocate resources or prioritize patients based on risk assessments might rely on historical data that reflect existing biases in healthcare access. For instance, patients from lower-income neighbourhoods or marginalized communities might receive lower priority because the system correlates socioeconomic factors with lower health outcomes, rather than considering the structural reasons behind these disparities.

As seen from the examples above, the deployment of AI systems in healthcare can contribute to systemic inequalities, even if the models powering those systems are technically well-designed. Such an outcome runs counter to the GDPR's overall goal of ensuring the protection of the rights and freedoms of individuals, such as their right to health or their right to privacy (which is affected by the large-scale accumulation of data about their healthcare). It can pose problems from an accuracy perspective, and it might also create issues from the perspective of non-discrimination law. Finally, some AI

applications might be subject to the AI Act's high-risk risks, in special if they are highly regulated under the Medical Devices Regulation. Hence, compliance with the requirements of data protection law will help **InnovaHospital** ensure that its embrace of innovative technologies will not come at the expense of the hospital's commitment to equitable patient care.

Session 4.3. Opacity as a risk

> By the end of this session, learners will be able to **describe** technical and non-technical sources of opacity surrounding AI systems. They will also be able to **estimate** how that opacity can create problems for compliance with data protection requirements.

A well-known problem with AI systems is their **opacity**. The expression "black box" has entered public discourse as a way to describe how the inner workings of AI systems and models remain hidden from the sight of the general public. However, the technical complexities we have discussed in Unit 2 often mean that even the organizations deploying AI systems might lack access to the information they need to make sense of how those systems work. In this session, we will examine potential sources of opacity and discuss their implications for organizations under data protection law.

In short, the opacity of AI systems can stem from technical or legal factors. Both tend to appear in practice, combining themselves to hide information from regulators, the general public, and data controllers themselves. This can be a legal problem in itself, to the extent that it prevents controllers from complying with their transparency and accountability duties. But it can also be a complicating factor in the various issues we discussed above, amplifying harm by making sure that organizations and data protection authorities cannot discover in time what is going on. It is not possible or desirable to eliminate those sources of opacity. Still, their potential impact on the rights, liberties, and interests of those affected by AI means that the legitimate grounds for opacity must be balanced with those other interests at stake. Therefore, organizations will need to adopt measures to deal with opacity in the AI technologies they create or use.

Two kinds of AI opacity

AI systems are often characterized by a high degree of opacity, which can arise from many factors. On the technical side, many AI models, particularly those based on deep learning, are complex and difficult to interpret. However, even when it is technically possible to make sense of an AI-based technology, other factors might be an obstacle to that. In particular, opacity can also be produced by the law. For example, some provisions in the German tax code prevent the disclosure of information about the

algorithms used by tax authorities for estimating fraud risk (Hadwick and Lan 2021). The interplay between **technical and non-technical factors** can contribute to our lack of understanding about what happens within an AI system or model.

When we think about the black box of AI, we often think about its technical complexity. AI models rely on intricate mathematical operations and advanced computational techniques. To understand these models—let alone to tinker with their inner workings—one must have specialized training. Even though recent developments in AI technologies reduce the specialized knowledge needed for using them, their components, such as the neural networks powering many AI models, remain inaccessible to non-expert users (Kolkman 2022). Experts might also struggle to make sense of those models, given the vast number of parameters and the complex architectures in which their components are arranged (Burrell 2016). As a result, making sense of what an AI system is doing is a task that can require considerable technical work.

However, as discussed above, that task is not always in the best interest of some actors. For example, government authorities might be unwilling to release some information about how their AI systems work, fearing that citizens might "game" the system to avoid detection. Or the providers of AI technologies might not want to release information about how they configure their AI models in order to prevent competitors from using that information to create better models.

The law recognizes various legitimate reasons why one might want to pursue secrecy, such as:

1. **State secrecy**, that is, the protection of information related to vital public functions.
2. **Trade secrecy**, that is, the protection of information related to how a business operates.
3. **Intellectual property law**, which might be used by organizations to deny access to the technicalities of an AI system or model.
4. **Data protection law** itself can be an obstacle to disclosure, for example if an organization argues it cannot disclose the training data for an AI model because it contains information about identified or identifiable natural persons.

Many of those legal grounds have been used to deny organizations access to information about AI.

Often, the denial of information is mostly directed at the **general public**, as seen in the examples above. But some of the concerns driving organizations towards confidentiality can also apply towards **downstream providers**. For example, a company that sells a general-purpose AI model might be afraid that its consumers will clone its model and

become competitors. The result is that legal opacity is sometimes used against the organizations that create and use AI technologies. Those providers and deployers find themselves in the unenviable position of being potentially responsible for the outcomes produced by technologies they have little margin to understand or control. As we shall see now, this situation has legal implications.

What AI opacity means for data protection law

As mentioned in the introduction to this session, AI opacity can lead to two distinct but related issues. If organizations lack visibility of the inner workings of an AI system or model that they use, they might be unable to comply with any legal obligations requiring them to release information about those workings. Additionally, AI opacity might hinder the detection of other sources of harms within an AI system, delaying their detection and response. Both implications of opacity are relevant for data protection law.

Regarding the legal obligations that are directly affected by opacity, one can focus primarily on issues of transparency and accountability.

- Articles 13–15 GDPR establish that the data controller must be able to disclose some types of information to the data subject, such as information about whether and how the automated processing of their data is used for making decisions without human involvement.
- Article 24 GDPR further establishes that data controllers are responsible for demonstrating compliance with the requirements of data protection law. Given that some of those requirements concern the technical means used to process personal data, complying with this duty requires information about how the system is set up.

These duties apply in any use of AI that involves the processing of personal data,[11] regardless of the use of AI. This means that organizations deploying or developing AI systems cannot invoke technical complexity as an excuse to discharge their duties. Instead, they are expected to adopt technical and organizational measures that mitigate said opacity.

Such measures are also relevant for the detection of risks associated with the use of AI-based technologies. Two examples can illustrate how opacity might amplify such risks:

1. Consider a scenario in which **UNw** decides to use an AI system for grading and assessing student performance. If certain groups of students consistently receive lower scores due to biases in the model, the university may not realize this issue if the system's decision-making process is too opaque to audit. This lack of

[11] For more details, see Session 6.1 of this training module.

visibility can allow discriminatory patterns to persist, even if the university has no intention of unfair treatment.

2. For **DigiToys**, opacity in its AI-enhanced toys might prevent the company from identifying privacy issues. If parents express concerns about how the toys process and respond to children's voices, the company may struggle to offer clear explanations or assurances due to the proprietary nature of the algorithms involved. This lack of transparency can erode trust and lead to reputational harm, even if the AI system functions as designed and complies with other legal requirements.

To the extent that organizations are obliged to adopt legal measures to address those risks, as we discussed in this unit, opacity can be an obstacle to compliance. It can increase the time necessary for identifying that a risk exists and for understanding its likelihood and severity. For example, **UNw** might struggle to detect the biased algorithm because the terms of service from the algorithmic tool it uses do not allow access to inner system parameters. In that case, risks might only be noticed once they have manifested and harmed students.

Even after the risk is detected, opacity might mean that an organization does not know exactly how it can address a problem. For example, it might be the case that **UNw**'s tool actually has settings that would allow for safe processing, but the university does not know about those settings. Opacity is not a problem just for data subjects, but for the controllers processing their data as well.

Conclusion to Unit 4

In this unit, we have seen the importance of distinguishing between security and safety in the context of AI systems. A secure AI system might still be unsafe for use, either because of its technical properties or because of problems with the context in which an organization wants to use it. Conversely, an AI system that is safe in light of those factors might still cause harms to individuals and groups if its security is not adequate for its task. Therefore, organizations need to pay attention both to cybersecurity and to the safety of their AI technologies in order to comply with the GDPR's requirements.

Toward that goal, we can highlight the following **takeaways** from this unit:

- Safety risks from AI can appear from a variety of sources, including but not limited to:
    - Technical and organizational shortcomings that prevent an AI system from delivering the promised results.
    - Unlawful or otherwise unethical uses of AI technologies, which can sometimes be *more* harmful if the system operates correctly.

    - The opacity of AI systems, which can prevent compliance with disclosure requirements and prevent the detection of other safety hazards.
- Safety risks must be addressed by **proactive measures**, both to prevent their occurrence and to mitigate the hams from any incidents during operation.
- Risks can be addressed by technical measures (that is, by changes to the design of an AI system or model) and organizational measures (that is, changes to its operation context).
- **Some risks cannot be fully eliminated, only mitigated**.
    - Some technical risks might follow from essential properties of the technology. Others might be solvable in theory, but an adequate solution might be beyond the state of the art. Last but not least, fully eliminating some risks might be too expensive in practice.
    - Likewise, some organizational risks are inherent to a technology's intended purpose, the context in which its meant to operate, or general societal arrangements that cannot be changed just for the sake of safe AI use.
- Whenever that is the case, organizations must decide whether they can mitigate the risks enough to make the use/development of AI worthwhile. If not, they might want to abandon it.

By keeping in mind the points above, one can have a clearer picture of why safety matters and why it might be threatened by the use and development of AI technologies. The rest of this training module will show various measures and safeguards that can be adopted to detect and respond to potential safety risks.

Prompt for reflection

The three kinds of safety failures discussed in this unit are complex. They can emerge from many sources, and it can be hard to find out whose actions caused the ensuing harms. Discuss who should be held accountable in these situations: the developers, the deploying organizations, or both? How can data protection officers (DPOs) play a proactive role in identifying and mitigating such risks before they materialize?

References

Adrien Bibal and others, 'Legal Requirements on Explainability in Machine Learning' (2021) 29 Artificial Intelligence and Law 149.

Maja Brkan and Grégory Bonnet, 'Legal and Technical Feasibility of the GDPR's Quest for Explanation of Algorithmic Decisions: Of Black Boxes, White Boxes and Fata Morganas' (2020) 11 European Journal of Risk Regulation 18.

Jenna Burrell, 'How the Machine "Thinks": Understanding Opacity in Machine Learning Algorithms' (2016) 3 Big Data & Society 1.

Madalina Busuioc, Deirdre Curtin, and Marco Almada, 'Reclaiming Transparency: Contesting the Logics of Secrecy within the AI Act' (2023) 2 European Law Open 79.

Roel IJ Dobbe, 'System Safety and Artificial Intelligence' in Justin Bullock and others (eds), *Oxford Handbook on AI Governance* (Oxford University Press 2022).

Stefan Gaillard, Cyrus Mody and Willem Halffman, 'Overpromising in Science and Technology: An Evaluative Conceptualization' (2023) 32 TATuP - Zeitschrift für Technikfolgenabschätzung in Theorie und Praxis 60.

David Hadwick and Shimeng Lan, 'Lessons to Be Learned from the Dutch Childcare Allowance Scandal: A Comparative Review of Algorithmic Governance by Tax Administrations in the Netherlands, France and Germany' (2021) 13 World Tax Journal.

Dominik Herrmann and Henning Pridöhl, 'Basic Concepts and Models of Cybersecurity' in Markus Christensen and others (eds.), *The Ethics of Cybersecurity* (Springer 2020).

Hartmut Hirsch-Kreinsen, 'Artificial Intelligence: A "Promising Technology"' (2024) 39 AI & SOCIETY 1641.

Daan Kolkman, 'The (in)Credibility of Algorithmic Models to Non-Experts' (2022) 25 Information, Communication & Society 93.

Nancy G Leveson, *Engineering a Safer World: Systems Thinking Applied to Safety* (The MIT Press 2012).

Steve McConnell, *Software Estimation: Demystifying the Black Art* (Microsoft Press 2006).

Inioluwa Deborah Raji and others, 'The Fallacy of AI Functionality', *2022 ACM Conference on Fairness, Accountability, and Transparency* (ACM 2022).

Luke Stark and Jevan Hutson, 'Physiognomic Artificial Intelligence' (2022) 32 Fordham Intellectual Property, Media and Entertainment Law Journal 922.

Charlotte A Tschider, 'Legal Opacity: Artificial Intelligence's Sticky Wicket' (2021) 106 Iowa Law Review Online 126.

Part II: The Life Cycle of an AI System

By the end of this part, learners will be able to:

- **differentiate** the various stages of an AI system's life cycle and the technical and organization decisions that take place at each stage.
- **assess** data protection risks that can emerge because of those technical and organizational decisions.
- **sketch** an initial set of compliance measures for the legal requirements that apply at each stage of the life cycle.
- **illustrate** how issues that are not addressed at earlier stages of the system's life cycle can propagate to later stages; and
- **propose** technical and organizational practices that can mitigate the risks associated with each life cycle stage.

One of the main challenges of AI regulation is that it deals with a moving target. Technologies change all the time, sometimes radically. Deploying AI in your organization in 2024 requires different kinds of technical work than it required in 2019, which in turn are very different than what AI developers did in 2010. At the same time, the social contexts in which those technologies are used can change considerably, too. The widespread enthusiasm for large language models seen in 2022 and 2023, for instance, was somewhat tempered since then as society became increasingly aware of risks associated with those technologies. Hence, the measures that govern AI technologies cannot remain static but must adjust to those new realities.

Both the GDPR and the AI Act feature adaptation mechanisms. Under Article 25 GDPR, data controllers are required to address the risks created by processing "both at the time of the determination of the means to processing and at the time of the processing itself". In short, this obligation reinforces that data protection is not a "fire and forget" duty. While measures in the initial design of a system can be crucial for ensuring adequate protection, they are not enough: data protection must be ensured in each individual processing, too.

In the AI Act, this moving targeted is captured by the notion of the "life cycle" of AI systems and models. Article 9 AI Act, for instance, requires the providers of high-risk AI systems to manage risks throughout the entire life cycle of an AI system, in particular by ensuring that the system keeps adequate levels of accuracy, cybersecurity, and

robustness.[1] Likewise, the AI Act stipulates that harmonized technical standards must deal with an AI system's energy consumption throughout its life cycle.[2] Yet, the concept of "life cycle" itself is not given a formal definition in Article 3 AI Act.

Such a definition is left, instead, to technical sources. The idea of a software life cycle is well-established among software engineers,[3] who use the term as a shorthand for the various technical processes involved in constructing and maintaining a computer system until the end of its operation. To better visualize those processes, software engineers often rely on **life cycle models**, which divide those processes into a succession of stages. This approach will guide the present training module.

More specifically, the training module takes as its starting point the life cycle model proposed by the international standard ISO/IEC 5338:2023. Future updates of this standard, or alternative standard such as those issued by European Standardization Organizations, might lead to different arrangements of the technical processes related to AI. But viewing those processes in an organized way will be useful for anticipating issues and incorporating data protection responses into what an organization already does.

Part II of the training module begins with **Unit 5**, which discusses the inception stage of the life cycle of an AI system, that is, the strategic decisions that shape whether and how an organization will use AI-based software. **Unit 6** then discusses AI-specific concerns that emerge with the use of personal data in the design and development of an AI system, followed by a discussion in **Unit 7** about how to evaluate AI systems before and after deployment. After that discussion, **Unit 8** considers what organizations must do to lawfully deploy AI systems for specific tasks. Finally, **Unit 9** considers their continuous obligation to monitor whether and how an AI system is functioning.

[1] Article 15 AI Act.
[2] Article 40(2) AI Act.
[3] See, for example, Kneuper (2018).

Unit 5. The Inception of AI Technologies

By the end of this unit, learners will be able to **identify** various interventions that can be made during the inception stage of a life cycle. In particular, they will learn how to:

- **inventory** AI systems in their organization; and
- **classify** AI systems based on the risk associated with their application.

AI systems and models are technical artefacts. They are created by somebody's deliberate effort, usually as a tool for some purpose. As discussed in Session 1.2 of this training module, this purpose is what will determine the applicable legal framework for systems and models. However, there is a huge gap between defining a purpose and actually creating a system that has a plausible claim at achieving that purpose. For example, some of the AI technologies we take for granted nowadays, such as large-scale text generators, were conceivable for a long time, but only recently it become possible to implement them. If a purpose is achievable, it takes a considerable amount of technical work to create a system or model that can do it.

The **inception** stage of the AI life cycle represents the starting point of that work. It refers to that moment in time when an organization is about to begin an AI project. At that point in time, the organization must make several choices that will influence how it develops (or purchases) AI:

1. It must identify whether and how the new AI project would contribute to addressing organizational needs. Each of the organizations used as examples in this module has its own reasons for pursuing AI:
    a. **UNw** sees its AI products as a way to cope with a growth in the number of students that was not met by the recruitment of academic and administrative staff.
    b. **DigiToys** has AI as the core element that makes its toys unique.
    c. **InnovaHospital** wants to use AI to improve its medical services and internal processes.
2. Once those general needs are defined, they must be translated into more specific **requirements**. A software requirement is a measurable stipulation of a condition that the AI system or model must meet before being deployed. It can be:
    a. **Functional** if it refers to technical properties of the system (or model). For example, **InnovaHospital** might stipulate that an automated diagnosis tool can only be acceptable if it performs considerably better than human doctors.

   b. **Non-functional** if it refers to a property that is not related to the main function of a system (or model) but is nonetheless desirable. One such requirement is the data minimization principle enshrined in the GDPR.
3. Considering those assessments, it can make an initial decision about how to move forward with an AI-based solution: whether the technology will be developed in-house, contracted from an external provider, or some other arrangement.

Those requirements, in turn, can reflect the perspectives of various stakeholders, such as different units within a business, potential clients of an AI-based product, or the communities affected by a proposed AI solution. Software engineering disciplines have developed various techniques to identify relevant stakeholders and elicit requirements that are relevant for an application. Such techniques will not be examined in depth here, but the materials indicated in the references offer an introduction to them.

What this unit covers, instead, is the roles that data protection professionals can play during this stage of the life cycle. **Session 5.1** sketches out those roles by highlighting the interplay between the purpose of an AI system or model and its role in data processing. **Session 5.2** then looks at the challenges involved in creating and maintaining up to date an inventory of AI applications, which can be essential for evaluating compliance with data protection law and the AI Act. Finally, **Session 5.3** discusses how AI systems can be classified into the various risk categories defined in the AI Act.

Session 5.1. Data protection tasks in the inception stage

> By the end of this session, learners will be able to **illustrate** how data protection professionals can be actively involved during the inception stage of an AI life cycle.

The inception stage is a strategic process, which takes place long before any data is processed by an AI system or model. Even in the absence of any actual processing, data protection law still creates obligations. Article 25(1) GDPR, for instance, stipulates that risks to data protection principles must be addressed also during “the determination of the means for processing”. While the specific technical means are chosen at the next step of the life cycle,[1] this provision already creates some obligations at the inception of an AI system or model. Therefore, data protection professionals must already be active at this stage.

[1] See Unit 6 of this training module.

A data protection professional must consider the specific risks that can emerge at this stage. That is, they must consider whether decisions about the need to create an AI system (or model), its intended purpose, and its functional and non-functional requirements can have unacceptable implications at some later point. For example, the AI systems that **InnovaHospital** wants to build for medical diagnosis give origin to concerns about the proper handling of special categories of personal data. Addressing such issues at an early point might prevent an organization from committing to an AI system or model that will need to be changed or abandoned later.

Initial assessments of lawfulness for processing

One risk that data protection professionals can diagnose at the inception stage is that of unlawful processing. Even before a system or model is built, the purpose for which it is meant might already signal the need to look more closely at some aspects. Any system that **DigiToys** uses in its toys, for instance, will need to be built in a way that is compatible with the processing of the personal data of children.

The AI Act adds red lines to processing, as it prohibits the use of AI in some practices, but it also authorizes some kinds of processing of special categories of personal data to avoid biases in high-risk AI systems. Therefore, a data protection professional can tell an organization about the implications of how they frame an AI system (or model)'s purpose.

Mapping legal requirements

Once the lawfulness of the proposed use of AI has been identified, a data protection professional can help the organization make sense of the applicable legal requirements. Some of those requirements are specific to AI technologies: Article 10 AI Act establishes certain data management and quality obligations, while Article 22(3) GDPR stipulates safeguards that must be observed in cases of automated decision-making based on personal data. To address this kind of data protection requirement, the data protection professional must be familiar with the kinds of data required for some AI tasks, as further studied throughout this module.

Other obligations reflect, instead, the flows of personal data that are needed before an AI system can be used. Consider, for instance, a scenario in which an organization relies on an AI-as-a-service tool offered from outside the EU. It can only do so if it takes the necessary steps to ensure that any personal data transferred to that service provider follows the GDPR's requirements for transfers of data to personal countries. Here, the data protection professional will benefit from their expertise on established mechanisms, such as standard contractual clauses or procedures and adequacy decisions.

Contracting for AI

In particular, the data protection professional can offer valuable guidance when it comes to the assessment of the contracts between an organization and its AI providers. Many AI models (and even full-blown systems) are not developed in-house or bought as closed products but hired as services, as Unit 13 of this training module explores in more depth. If that service is acquired between relatively equal parties, an organization might have the flexibility to include some clauses that facilitate its own performance of data protection clauses. For example, **InnovaHospital** might require its non-EU contractors to follow standard contractual clauses as a safeguard for the processing of personal data.

Such negotiations are not always possible. Because many AI systems and models are developed by large organizations, those providers have considerable power when setting the terms of purchase. A recent study of large language models (Edwards et al. 2024), for instance, has found that most of them are offered through contracts of adhesion, in which buyers can only accept or reject the proposed clauses.

In this context, a data protection professional will need to support an organization in evaluating whether those clauses are compatible with the organization's own data protection duties. For example, if a provider denies to a deployer information about how its AI systems work, contracting with that provider might create a situation in which the deployer cannot comply with some of its duties, such as the information rights from Articles 13–15 GDPR.[2]

Taking stock of how AI is used

Finally, a data protection professional can help an organization keep track of the AI systems it already has. Each of these AI systems will raise its own data protection issues, and so an unnoticed AI system is a likely source of data protection exposure.

Additionally, a global vision of what is going on within organization is important for understanding data protection issues that might emerge from the interaction between different systems and models. For example, it might be the case that the combination between two separate AI systems that **DigiToys** uses to analyse personal data supplies enough information for identifying the owners of some toys.

Such risks must be addressed as part of the overall strategy for data protection compliance. However, as we will discuss in the next session, determining what AI systems are in use within an organization is not always a straightforward task.

[2] See Unit 11 of this training module.

Session 5.2. Mapping the uses of AI

By the end of this session, learners will be able to **formulate** approaches for identifying whether and how AI is used within their organization.

Any organization that develops or uses AI systems or models is subject to some legal obligations. In Part I of this training module, we saw that the shape of those obligations may vary with the type of technology in question and the context. Nonetheless, the organization is still bound to obligations regarding the AI system itself and any personal data it processes. Understanding whether AI is in fact in place is a necessary first step for discharging those obligations. Yet, it is entirely possible that an organization makes extensive use of AI without knowing it.

Consider two examples in which an organization benefits from the *background* use of AI.

1. The **UNw** university has purchased access to a workplace software suit supplied by a large company *OfficeCorp*. To preserve its dominance in the workplace software market, *OfficeCorp* is always pursuing new ways to boost user productivity. And, with the latest developments in AI technologies, it has decided to use some AI-powered tools in its background work. One of those tools, for instance, tries to optimize meeting schedules across teams in the company by suggesting the best timeslots and proposing agendas. It does so without requiring the end-users of the tool—in this case, **UNw** staff—to interact with an AI system. Hence, its use of AI might even go unnoticed.
2. The toy company **DigiToys** decides to outsource some of its data analysis to an external provider, *AnalyticsRUs*. From the contract between those two companies, *AnalyticsRUs* is obliged to follow best statistical practices and make sure that it uses data in accordance with data protection principles, such as minimization and (whenever possible) anonymization. That company, however, reserves the right to protect its trade secrets, and so it does not offer **DigiToys** direct access to information about whether and how AI is used for data analysis.

Both cases illustrate how an organization might not be aware of AI being used on its behalf. In Session 5.1, we saw that such ignorance is not an unavoidable consequence of relying on external providers: an organization can and should pursue information about AI used by third parties on its behalf. And, if an organization does not do so, that is not an excuse for sidestepping its legal duties under either the AI Act or the GDPR. Not knowing about the use of AI might prove to be a costly decision.

Ignorance about the use of AI can emerge even when AI systems and models are deliberately used within the organization. One reason for that is that innovation does not always start from the top. Some AI systems are not developed or used uniformly across an organization, appearing instead from the initiative of individuals or small teams. For example, a human resources professional at **InnovaHospital** might read a news article about ChatGPT and create a prototype tool that automatically summarizes the files of workers, reducing the number of documents that need to be read for designing their career path. Or the professors at **UNw**'s computer science department might decide to create a chatbot that answers questions from students of the introductory programming course, freeing up more time for research and grant applications. These so-called **shadow AI** initiatives reflect initial inspiration, but often remain undetected until they are successful—or lead to harms to an organization's interests. If individuals and groups fail to report the use of AI, or believe such uses can go unreported, it might be a long time before their existence come to the attention of a data protection professional.

Other aspects of an organization's culture might contribute to lack of visibility about the use of AI systems. At **UNw**, for instance, there is a massive rivalry between the departments of computer science and electrical engineering, both of which carry out research on AI. So, if one of those departments decides to create an AI system for its own purposes, as in the example above, it is unlikely to involve the other, and it might even try to keep the use of the system secret. Such secrecy can affect an organization in several ways, ranging from the waste of effort in creating (or buying) systems that already exist elsewhere in the organization to a lack of proper supervision of systems that go unnoticed. Therefore, knowing what is going on within an organization is an important starting point for the inception of any AI tools.

Reasons for maintaining an AI inventory

The most straightforward approach to this issue is to create an **inventory** of AI systems that exist within an organization. A data protection professional needs to know about AI-based processing to oversee any processing operations. If they organize that information into a structured form, such as a list or an internal database of AI uses within an organization, they will then gain a comprehensive view of what AI technologies are in use and how they interact with one another.

At the end of the day, each organization is free to organize this information in any way they want. It might not have a centralized list, or it might not store much information about each AI system. Still, an organization would benefit from having easy access to the information they need to comply with existing legal requirements. Otherwise, it might need to gather that information each time they need to demonstrate legal compliance with a data protection or AI Act requirement. This session does not provide a comprehensive set of information that should be kept about each AI system or model. It focuses, instead, on the task of listing those systems and models in the first place.

Such a list, in itself, would already facilitate conformity with some legal obligations. Awareness that AI is being used for a particular application contributes to the AI literacy that every organization providing or deploying AI must foster under Article 4 AI Act. Providers of high-risk AI systems listed in Annex III AI Act are obliged to register those systems before they are placed on the market or put into service,[3] an obligation that falls on the deployer of some public-sector applications.[4] Furthermore, awareness of the use of AI is needed if data subjects decide to exercise their rights to obtain meaningful information about how a decision involving an AI system or model is produced.[5] This list is not meant to exhaust the duties in which information about the use of AI is relevant, just to exemplify why an inventory can be useful for overall compliance.

Building the AI inventory

But what can a data protection professional do to build or update an inventory of AI systems? Comprehensive coverage of all AI systems in an organization will require a good strategy for collecting information, as well as a constant effort to ensure the inventory stays up to date. There is no alternative to a thorough examination of data practices to ensure nothing is being missed. But the following paragraphs highlight a few issues that one should not overlook.

Communication as a source of information about AI systems

First, **communication** is essential to ensure that no AI system or model is being overlooked. By gathering information from departments within the organization, as well as from individuals, the data protection professional can find systems and models that would otherwise escape their attention. They can also establish themselves as a reference point for AI within the organization, thus creating a virtuous circle in which individuals and departments proactively supply information about AI.

Consider a few forms of communication that might be relevant:

1. The data protection professional can make **direct requests** of information to departments about any AI projects that they might be using. This approach is likely to yield information about systems that a department is developing, or that it is aware of using. But a data protection professional might need to assist the department with guidance to find AI systems being used in the background.
2. A good rapport between the data protection professional and the organization's inner structures might lead to **voluntary reporting** of information about AI. For example, the **UNw** data protection officer might

[3] Article 49 AI Act.
[4] Article 26(8) AI Act.
[5] Articles 13–15 GDPR.

position themselves as a neutral observer and thus establish contact with both rival departments.

3. Data protection professionals might gain valuable information from addressing **individual queries**. For example, a human resources specialist at **InnovaHospital** might contact the hospital's DPO if they are not sure that the new tool they are using relies on AI. After investigation, the data protection professional might find out that AI is indeed being used. Even if it is not, this investigation is likely to uncover relevant data processing that needs to be addressed.

Classification uncertainties

Second, the data protection professional's habitual diligence is particularly relevant given the various **uncertainties** about AI mapped in Part I of this training module. Despite the technical and social complexities of AI systems, the information obtained through communication processes cannot be taken at face value. Instead, the professional must carry out a dialogue with the technical experts within the organization (and at the external providers) to evaluate the legal implications of the systems being analysed.

For example, a department within an organization might say that their new tool is an AI system in order to secure extra funding or to gain prestige from using advanced technologies. But, upon further inspection, their system might not meet the AI Act's criteria for an AI system.[6] Or, conversely, an individual software developer within an organization might want to avoid mentioning their use of AI in a project in order to avoid having to deal with internal compliance requirements before a project is mature enough.

So, **any information about a supposed AI system must be thoroughly verified before it is added to the inventory.**

Keeping the inventory up to date

Third, the inventory must be **updated** often:

1. Even if an organization does not develop its own AI systems (or develops just a few of them), its data processors might decide to use AI for some reason. In the **UNw** example, *OfficeCorp*'s business decision to use AI was taken without prior notification to the university, which would only find out about it after the new AI systems were implemented.
2. Some AI systems that were originally listed can be deactivated.

[6] See Unit 1 of this training module.

3. The legal classification of an AI system under the AI Act might change over time. For example, Article 7 AI Act allows the European Commission to update the list of high-risk AI systems.

An inventory that is not verified from time to time might lose its usefulness or even become a misleading guide to AI in an organization.

Public or private inventories?

Fourth and finally, the organization must decide how much it wants to **disclose** the inventory. If an inventory is made accessible within the organization (or even to the public), departments and individuals might feel more inclined to supply information and keep it updated. Additionally, the availability of this information can generate value for businesses. As an example, the Brazilian judiciary has a list of AI models developed by each court, allowing other courts to benefit from tried-and-true solutions rather than creating their own system from scratch. Those benefits of transparency, however, might be outweighed by other organizational factors, such as the need to preserve trade secrets.

Session 5.3. The purposes of AI technologies

> By the end of this session, learners will be able to **assign** different AI systems to the various legal frameworks established by the AI Act.

Once a data professional knows what AI systems and models are used in an organization, they can start assessing the organization's data processing architecture. Each system or model is used for one or more tasks, which often involve distinct kinds of personal data. Additionally, those systems are often **interconnected**, in the sense that the output of one system can act as the input for another. For example, **InnovaHospital** might use an AI-based solution to analyse the data generated by the interactions between patients and the hospital's chatbots. It is necessary, therefore, to have a sharp vision of the function and the interactions between existing AI systems and models.

Here, it is important to distinguish between two kinds of purpose that are relevant here. On the one hand, each **AI system or model** is designed for a purpose, that is, for carrying out one or more specific tasks. The system or model's purpose is relevant for determining the rules applicable under the AI Act. On the other hand, those systems and models might also process personal data, either during their training process or during their operation. As such, the purposes of each **processing** operation become relevant for the application of data protection requirements to the system. It is now time to examine those two kinds of purpose. The purposes of data processing will be

examined more closely in other units of this training module. Now it is time to look more closely at the legal classification of purposes for AI systems and models.

Models and systems are subject to different legal frameworks under the AI Act. However, as discussed in Session 1.2 of this training module, all those frameworks are organized around the purpose for an AI model or system. General-purpose models—which can be used to power systems designed for various purposes—are subject to special rules, some of which apply to every general-purpose. If a model can only be used for a specific purpose, then it is not directly regulated. Still, conformity with the rules that apply to an AI system in light of its purpose might require changes to the model powering it. So, one must determine *why* an AI system is being used in order to find out the applicable rules.

Prohibited AI applications

Article 5(1) AI Act establishes a list of prohibited AI practices. In some cases, those practices are themselves illegal or at least questionable, but the addition of AI would have the potential to amplify the harms. One such prohibition is that of Article 5(1)(a), which bans the use of AI to materially distort the behaviour of a person or group of persons in a way that causes or is likely to cause them to harm themselves or others. This means, for instance, that an application that uses AI to incite conflicts within a society is likely to be unlawful.

Other prohibitions deal with practices that are not in themselves unlawful but become risky when done at the scale enabled by AI. For example, the AI Act bans the use of emotion inference systems in the workplace or in educational institutions.[7] This provision also illustrates another feature of the AI Act's system of prohibitions: they sometimes allow for exceptions. In this case, the use of emotion inference systems is allowed when the system is intended to be put in place or into the market for medical or safety reasons. Those exceptions can be quite big, as shown by the fact that most of Article 5 AI Act is dedicated to laying down conditions in which law enforcement can use real-time biometric identification systems that are theoretically prohibited by Article 5(1)(h). But, without such an exception, no number of technical safeguards can allow the lawful use of a system covered by Article 5.

High-risk AI systems

The AI Act distinguishes between two types of high-risk AI systems. Some systems are classified as such because they are used in products that are, in themselves, subject to **harmonized product safety** law at the EU level. Others are classified as such because the EU lawmaker has deemed that the **risks they create to fundamental rights, democracy, and the rule of law** are big enough to warrant special attention. For the

[7] Article 5(1)(f) AI Act.

most part, those two kinds of high-risk AI systems are largely subject to the same rules. Still, there are a few differences between the classification procedures applicable to each one.

Product safety law

When it comes to systems already covered by product safety law, Article 6(1) AI Act establishes two conditions for the high-risk classification.

The first condition relates to the intended purpose of the system: the system must be a product covered by the product safety legislation listed in Annex I AI Act, or a safety component of such a product. For example, the toys produced by **DigiToys** are regulated by the Toy Safety Directive,[8] and so they meet this first requirement. Likewise, if **InnovaHospital** decides to incorporate AI into medical devices, those devices might be covered by the existing regulations on medical and *in vitro* diagnostic devices.[9] So, they would meet this first criterion.

The second condition for the application of rules on high-risk AI comes from product safety law itself. Under Article 6(1)(b) AI Act, an AI system is classified as high-risk if the product in which it is used must undergo a third-party conformity assessment before being placed on the market:

1. In the case of **DigiToys**, the company has decided that such an assessment is necessary due to the nature of their products and the lack of technical standards applicable to smart toys,[10] which means they also become subject to the AI Act's rules.
2. In the case of **InnovaHospital**, classification is more contextual, as the applicable regulations have a complex mechanism for determining which applications require third-party assessments.

But, whenever a device without AI would need such an assessment, the use of AI means the rules on high-risk systems become applicable.

Risks to public values

Many of the high-risk applications covered by the AI Act have no precedent in product safety law. Article 6 AI Act stipulates that all systems listed in Annex III AI Act are high-risk, unless they are covered by one of the derogations present in Article 6(3) AI Act. Scholars and civil society organizations have pointed out that there is no underlying logic to this list of high-risk applications. It reflects, instead, applications that the EU

[8] Directive 2009/48/EC.
[9] Regulations (EU) 2017/745 and 2017/746.
[10] Article 19 Directive 2009/48/EC.

lawmaker saw as creating particular risks to public values, in particular the protection of fundamental rights, democracy, and the rule of law.

The AI Act's recitals offer guidance about the risks that the lawmaker associated with some—but not all—of the applications listed as high-risk. But, for the most part, the determination of which risks apply in each context is left to the deployers and providers of those systems.[11] Therefore, those actors need to know whether they systems are covered by one of the various rubrics of Annex III AI Act.

AI systems posing a high risk to fundamental rights

Annex III to the Act indicates eight types of application considered as high-risk:

1. Biometrics, in so far as the use of AI is permitted under EU or national law.
2. Operation and management of critical infrastructure.
3. Education and vocational training.
4. Employment, workers management, and access to self-employment.
5. Access to and enjoyment of essential private services and essential public services and benefits.
6. Law enforcement, in so far as the use of AI is permitted under EU or national law.
7. Migration, asylum, and border control management, in so far as the use of AI is permitted under EU or national law; and
8. Administration of justice and democratic processes.

Within each point, the EU lawmaker has designated one or more applications considered high-risk. Any other applications within that domain are not classified as such. For example, the use of AI for risk assessment in pricing in relation to natural persons for life and health insurance is covered under Point 5 above. By exclusion, risk assessment and pricing with regard to the insurance of legal persons or of other types of insurance for natural persons is exempt from the rules on high-risk AI. So, the rules on high-risk AI under Annex III only apply to a narrow set of applications designated as especially risky.

Derogations from the high-risk classification

A system might escape the rules on high-risk AI even if it is designed for a listed purpose. Article 6(3) AI Act stipulates that a system is not considered high-risk if it does not "pose a significant risk of harm to the health, safety or fundamental rights of natural persons". In general lines, this derogation covers situations in which the AI system plays only a marginal role in the outcomes. As currently formulated, this is the case whenever one of the four conditions below applies:

[11] Articles 26 and 9 AI Act, respectively.

a. **The system is intended to perform a narrow procedural task**, such as automatically archiving the work done by a human.
b. **It is intended to improve the result of a previously completed human activity**, for example by improving a report written by a human analyst.
c. **It is intended to detect decision-making patterns or deviations from prior decision-making, without replacing or influencing human assessment**, for example by flagging whether a transaction involves a much larger monetary value than usual; or
d. **The AI system is intended to perform a preparatory task to an assessment relevant for the purposes listed in Annex III**, leaving the assessment itself in human hands.

The examples listed above are merely indicative, as the degree of human involvement in a particular case might mean that the derogation is not applicable. Determining its applicability requires a contextual assessment. The only general rule in that regard is that any **system that carries out profiling in the context of Annex III applications is considered high-risk**, regardless of any subsequent human involvement.

The application of those derogations falls entirely to providers. Under Article 6(4) AI Act, a provider must evaluate whether their system is covered by one of the derogations above. If they consider that is the case, they must document their assessment before the system is put on the market or placed in the service. That documentation can be requested by national competent authorities at a later stage, who might question the provider's assessment. Until and unless an authority does so, the provider is obliged to register the system in an EU-wide database for high-risk AI systems[12] but is not subject to any other of the obligations surrounding high-risk systems.

Specific rules for specific purposes

The purpose of an AI system is also relevant for determining whether some of the AI Act's special rules are applicable. When it comes to high-risk AI, Article 27 AI Act stipulates that some deployers of those systems must carry out a fundamental rights impact assessment of the deployed system. This obligation covers all deployers that are governed by public law, or that are private entities providing public services. Furthermore, it also applies to deployers using AI systems for evaluating creditworthiness of natural persons (including by credit scoring) and for risk assessment and pricing for life and health insurance. Therefore, the high-risk legal framework can suffer some adjustments depending on the purpose of the system.

Furthermore, Article 50 AI Act establishes some requirements that AI systems must observe regardless of their risk classification:

[12] Article 49(2) AI Act.

1. Providers of systems meant to interact directly with natural persons must design and develop the system in a way that allows natural persons to be informed that they are interacting with AI.
2. Providers of AI systems generating synthetic audio, image, video, or text content must ensure that the outputs are market in a machine-readable format and detectable as being generated or manipulated by AI.
3. Deployers of an emotion recognition or a biometric categorization system must inform the natural persons exposed to that system of its operation; and
4. Deployers of an AI system that generates or manipulate image, audio, or video content constituting a deep fake must disclose that the content has been artificially generated or manipulated.

All those rules admit exceptions, which must be analysed on a contextual basis.

Conclusion to Unit 5

The safe use of AI requires close attention to the purposes for which AI systems and models are proposed. If that purpose is in itself unlawful, no amount of software engineering can make it acceptable in the eyes of the law. But even if the purpose is lawful at a first glance, different purposes raise different risks to data protection principles and other values protected by the law. As such, the purpose for which a system (or model) is built or used affects the legal duties to which a provider or deployer must comply.

Some of the risks created by AI might be anticipated early on. To do so, the previous sessions have highlighted a few good practices:

1. **Creating and keeping updated an inventory** of AI systems within an organization.
2. **Involving various stakeholders** within an organization in the process of keeping that inventory up to date.
3. Evaluating the **contractual terms** under which popular AI tools are offered.
4. Helping business stakeholders **identify potential implications** for data protection of the requirements they are mapping for an AI system.
5. **Using the inventory to evaluate** whether a system interferes with (or is interfered by) other systems within the organization; and
6. **Assessing the purposes** of AI systems before they move on to the development stage.

After completing those preliminary tasks, and organization should have a much clearer picture of what they use AI for and what needs a new system would address. This suggests that involving a data protection professional at the earliest stages of AI development can help organizations avoid not only legal liability for breaches of the law

but also the duplicated work that would come from redundant or even unlawful AI projects.

For the data protection professional, this early assessment process also contributes with visibility of the AI systems and models that will process personal data within the organization. Furthermore, classifying those systems in accordance with the AI Act is important to understand whether and how the general data protection obligations are modified by any AI-specific obligations. In the next stages of the AI life cycle, we will focus on the various data protection obligations that emerge throughout the life cycle.

Prompt for reflection

The chapter emphasizes the creation and maintenance of AI inventories to ensure compliance and accountability. Why do you think organizations might struggle with this task, and what practical strategies can a data protection officer implement to overcome these challenges? Consider how organizational culture, such as the rivalry at **UNw** or the reliance on third-party providers like **DigiToys** (or examples from your own organization!), might impact this process.

References

‘ISO/IEC/IEEE International Standard - Systems and Software Engineering – Software Life Cycle Processes’ [2017] ISO/IEC/IEEE 12207:2017(E) 1.

Marco Almada and Nicolas Petit, ‘The EU AI Act: Between the Rock of Product Safety and the Hard Place of Fundamental Rights’ (2025) 62 Common Market Law Review.

Rashidah Kasauli and others, ‘Requirements Engineering Challenges and Practices in Large-Scale Agile System Development’ (2021) 172 Journal of Systems and Software 110851.

F Khomh and others, ‘Software Engineering for Machine-Learning Applications: The Road Ahead’ (2018) 35 IEEE Software 81.

Ralf Kneuper, *Software Processes and Life Cycle Models: An Introduction to Modelling, Using and Managing Agile, Plan-Driven and Hybrid Processes* (Springer International Publishing 2018).

David Lehr and Paul Ohm, ‘Playing with the Data: What Legal Scholars Should Learn About Machine Learning’ (2017) 51 UCDL Rev 653.

Silverio Martínez-Fernández and others, ‘Software Engineering for AI-Based Systems: A Survey’ (2022) 31 ACM Trans Softw Eng Methodol 1.

Mariana Maia Peixoto, ‘Privacy Requirements Engineering in Agile Software Development: A Specification Method’ in (CEUR-WS 2020) Joint Proceedings of REFSQ-2020.

Luke Stark and Jevan Hutson, ‘Physiognomic Artificial Intelligence’ (2022) 32 Fordham Intellectual Property, Media and Entertainment Law Journal 922.

Rob van der Veer, ‘ISO/IEC 5338: Get to know the global standard on AI systems’ *Software Improvement Group*. Accessed 26 September 2024.

Titus Winters and others (eds), *Software Engineering at Google. Lessons Learned from Programming over Time* (O’Reilly 2020).

Unit 6. Designing and Developing AI Technologies

By the end of this unit, learners will be able to **assess** how various kinds of decisions by software developers and by the organizational stakeholders commissioning an AI system affect its use of personal data.

Once an organization has an initial idea of what it expects AI to do, it can start the work of building (or acquiring) the technologies needed for that purpose. To do so, the organization must make decisions about the various technical components of an AI system. What components will be used to build this system? How do these components connect to one another? How will they be integrated into existing computer systems within an organization? What data will be used to train the model powering the system? What data will be used in its day-to-day operation? Those choices are just a few of the technical decisions that impact how an AI system or model processes personal data.

At this point, it is important to distinguish between two kinds of technical decisions that are relevant from a data protection perspective.

Data processing within the AI training process

Some technical decisions at this stage result in the actual processing of personal data. For example, the **UNw** university might decide that it needs to use data about individual students to create a model that can forecast their risk of failure in difficult courses (to propose support measures to those students). Any processing of personal data during the training process, just like in any other moment, remains in principle covered by data protection law.

Not all kinds of personal data processing, however, are covered by EU data protection law. Article 2(2) GDPR lists four kinds of processing that lie outside the regulation's scope:

1. **In the course of an activity which falls outside the scope of Union law***:* this carve-out is unlikely to apply to AI systems processing personal data. Since the AI Act lays down rules on how AI systems are placed on the market, put into service, or used within the EU, any systems covered by it are within the scope of EU law.
2. **By the Member States carrying out activities within the scope of the EU's Common Foreign and Security Policy**. This exception will not apply to most public or private uses of AI, either.
3. **By a natural person in the course of a purely personal or household activity**, an exception that must be construed narrowly (see, for example, Papakonstantinou and de Hert 2023).

4. **By competent authorities in the criminal law contexts covered by Directive (EU) 2016/680,** which itself offers a set of data protection safeguards.

Any other processing of personal data is covered by the GDPR.[1] So, the application of its provisions can only be avoided by training a system solely on non-personal data, a possibility we discuss in Session 6.1 of this training module.

Determining the means for future data processing

The second kind of relevant technical decision pertains to technical decisions that will affect how the AI system or model will process personal data once it is placed into service or otherwise used. Those decisions stipulate certain aspects of the system's functioning, such as:

- The training algorithm that will be used to create an AI model.
- The training, test, and validation datasets that will be processed by that algorithm.
- The metrics that will be used to evaluate the training process (see Unit 7).
- The software libraries that will be used to implement the model or system.
- The choice of the input parameters that will be given to an AI system; or
- The interfaces between the AI systems and other systems operated by an organization.

All of those are **choices**. It is rarely the case that any of those technical problems can only be solved in a single way, which means that two systems (or models) created in response to the same requirements can have vastly different technical arrangements. But one thing these choices have in common is that none of them is solely responsible for the processing of personal data.

Still, they shape how an AI system or model functions. Different technical arrangements will process data in diverse ways, and lead to different outcomes. Consider a situation where **DigiToys** can choose between two systems that allow their toys to interact with children. One of them allows for smoother interaction, but it demands that the toy collect considerable amounts of data and is prone to occasional errors. The other affords a more limited set of interactions with children but needs less data and does not create as many errors, while still being more interactive than the competitor's toys. The choice between those two options will affect how much data **DigiToys**'s products will process in the future.

Still, those future-looking decisions remain covered by the GDPR. Under Article 25(1) GDPR, data controllers are required to address the risks stemming from processing "at

[1] Except for EU institutions, bodies, offices, and agencies, which are covered by Regulation (EU) 2018/1725.

the time of the determination of the means for processing", not just when it takes place.[2] If the AI system or model falls within the scope of the AI Act's rules for high-risk AI systems or general-purpose AI models, there are additional rules that must be observed before a system can be placed on the market, put into service or used. Legal compliance is not a matter for the moment when an AI system finally processes data, but something that must be considered throughout the entire life cycle of any system or model.

The software development process

Data protection professionals can face various difficulties in evaluating the decisions made at this stage of the AI life cycle. Some of them relate to the technical complexity of the development of AI systems and models. The topics on Part I of this training module are geared towards allowing collaboration with technical experts, but they do not capture the full technical nuance of all those topics. Hence, it is necessary to maintain an ongoing dialogue with software developers and engineers within an organization.

Further difficulties come from the fact that the design and development process can take various forms:

1. In agile software development, systems and models are developed iteratively. Starting from an initial idea of what the technical product should do, the technical team creates a first version, which is then refined with additional development work. In this process, both the system and the technical requirements change as time goes by, and there is a tendency to avoid formal documentation of decisions.
2. In waterfall software development, requirements are exhaustively defined at the beginning of the life cycle. Once that is done, the development process follows a linear sequence of stages: programming only begins after all requirements have been defined, the software is tested only after everything has been programmed, and deployment only happens when a system has been fully tested.

Most AI systems and models are developed somewhere in-between one of those two development models, including elements from agile practices and more traditional development modules.[3] As such, any list of technical decisions to be monitored would likely include some steps that are not followed in practice within a given organization or omit relevant development practices.

To illustrate the kind of relevant practices that a data professional must attend to, this unit focuses instead on three processes that are likely to take place within most

[2] For more on this topic, see Unit 13 of this course, as well as (Almada et al. 2023).
[3] Additionally, safety-critical systems such as those used in the aviation sector are often subject to particularly strict practices in their development process. Analysing those practices goes beyond the scope of the present training module.

organizations. **Session 6.1** discusses how the developers and designers of AI systems are classified under the GDPR and the AI Act, as that classification will affect the duties that apply to them. **Session 6.2** details those duties with regards to the acquisition of data for the development process. **Session 6.3** then outlines how the data protection principles of the GDPR can be applied in the AI development process, as well as pointing out the rules that apply to the processing of personal data in that process.

Session 6.1. The legal roles of AI developers

By the end of this session, learners will be able to **distinguish** between forms of software development involved in the creation of an AI system and **classify** those providers under the GDPR and the AI Act.

Once an organization decides it needs an AI system, it can do obtain one in a few ways:

1. It can **develop the system in-house**, creating a solution tailored to its own needs. For example, **InnovaHospital** might use its extensive collection of radiological data to create a system that automates the reading of scans for certain diseases.
2. Alternatively, the organization might decide its needs can be addressed by **technologies available on the market**:
    a. By **fine-tuning** those tools. For example, the professors at **UNw** might decide that they can create an automated system for answering student questions by starting from ChatGPT and doing some extra training to fine-tune it to the specific topics of the courses they teach.
    b. By **integrating ready-made systems** into their existing infrastructure. For example, **DigiToys** might license the use of a data analytics system to process all the data it collects from the toys, connecting that system to its databases via an application programming interface (API).
3. Or it might **procure** the entire system from outside sources.

The first two items all entail that an organization is doing some form of software development. Still, the software development work done under each item requires several types of personal data use and of technical skills. Going back to the examples above:

1. **InnovaHospital** will need to ensure it has software developers that can handle the construction of an AI model from the potential training data, as well as the integration of that model into the system. It will also need to determine whether the data it uses meets the criteria for personal data, and, if so, comply with them.
2. For the solutions based on ready-made components*:*

a. The professors at **UNw** will need to collect the data that is relevant for their application and figure out how to carry out the additional training on ChatGPT, a process that is simpler and less expensive than training an entire large language model. They will also need to evaluate compliance with personal data requirements.
b. **DigiToys** will not need to do any AI-specific software development. Still, it must evaluate whether the programming it does to connect the AI system with their existing systems processes personal data. For example, it might be the case that the system receives personal data for its operations.

To the extent that the data created, used, or otherwise processed during those processes relates to an identified or identifiable natural person, it will qualify as personal data. Likewise, the technical decisions made during those development processes become relevant to data protection law to the extent that the ensuing AI systems or models store or otherwise process personal data. Hence, the organizations developing and designing AI technologies have obligations regarding the processing of personal data during the training process.

Under both the GDPR and the AI Act, an organization's obligations depend on the role it plays in processing. Within the AI Act, classification is relatively straightforward. Anyone who develops an AI system or model is a provider,[4] unless one of the exceptions in Article 25 AI Act applies. Likewise, anyone using an AI system under their own capacity qualifies as a deployer, except in the case of personal non-professional use.[5] Classification within the GDPR regime is slightly more nuanced.

AI developers as data processors

If an organization is developing an AI system or model for its own, internal use, classification is straightforward. From a data protection perspective, the organization meets the definition of a data controller[6] both regarding present and future processing:

- The developer is the one determining why, when, and how personal data will be processed during the training.
- The technical choices it makes will determine the means through which the AI system will process in the future.

Classification under the GDPR becomes more complex when an organization develops an AI system or model intended for the use of others. During the training stage, the role of the developer will depend on the degree of independence of its actions. If the buyer provides detailed instruction on how the developer must conduct any data processing

[4] Article 3(3) AI Act.
[5] Article 3(4) AI Act.
[6] Article 4(7) GDPR.

during the training process, the developer organization becomes more of an executor of the buyer's will than an independent controller of the processing in training. Conversely, the responsibility of the developer grows in accordance with the amount of discretion it is afforded when it comes to determining the means for processing.

Suppose **DigiToys** decides to hire **InnovaHospital** to create an AI system that can diagnose respiratory illnesses in children, which will be incorporated into a new line of toys:

- Given the expertise of each organization, the toy company might decide to adopt a hands-off approach and leave the hospital free to choose what kinds of data processing are needed to train the model. In that case, **InnovaHospital** is still the **controller** of that processing from a legal perspective.
- It might be the case, instead, that **DigiToys** decides to provide strict instructions on whether and how the hospital is to process personal data. For example, the contract between the two might supply detailed stipulations of what is to be done during the development process. If those stipulations meet the requirements of Article 28(3) GDPR, **InnovaHospital**'s discretion is extremely limited. Hence, control of processing rest with the toy company, and the hospital is merely a **processor**.
- Many cases fall in-between those two extremes. For example, **InnovaHospital** might have considerable liberty to make its technical choices but rely on some data provided by **DigiToys**. Or both organizations might collaborate in determining the technical specifications of the system's data and algorithms. In such cases, a data protection professional needs to check whether the situation amounts to **joint controllership** of the processing.

Responsibility for subsequent processing

If an AI system or model is created for the use of others, its developer might be tempted to think they have no obligations regarding this subsequent use. After all, their system or model is just the technical means used by somebody else to process personal data, and it is this other who determines the means and purposes of processing. However, there are some circumstances in which a developer might have a role in the use of the AI system:

- **Joint controllership might emerge** if the developer is also involved in determining the purposes for processing. For example, if **DigiToys** and **InnovaHospital** are both involved in the decision of adding the diagnostic medical tool for the toy, then the hospital is involved in the determination of the means (because of its role in determining the technical arrangements of the AI system) and the purposes of processing, thus meeting the elements of controllership.

- **The developer might instead be a processor** for those subsequent instances of processing. For example, it is common nowadays to see AI-as-a-service arrangements, in which the buyer acquires access to an AI-powered tool on a subscription or pay-per-use basis instead of having to run their own system.

Both cases make developers potentially responsible, to a lesser or greater extent, for harmful outcomes stemming from the use of the AI system or model they provide. Therefore, a data protection professional cannot take for granted that the developer is entirely detached from any subsequent processing from their AI system.

Dividing responsibilities between developers and (other) controllers

In situations of external processors, or even of joint controllership, it is necessary to clarify how responsibilities are divided. Under Article 26 GDPR, joint controllers are required to come to an arrangement between them on how to assign those responsibilities, unless such an assignment is made by EU or national law. Similarly, Article 28(3) GDPR provides a quite extensive list of elements that must be present in the contract between a data controller and a data processor. Those elements remain unchanged when it comes to the relationship between an AI developer and downstream actors relying on their products.

Yet, one must be aware of the strong asymmetry that exists between developers and buyers in particular contexts. Some of the most advanced AI technologies that exist today, such as the large language models discussed in Unit 13, require massive amounts of data and computing resources for their construction. As such, the state of the art is concentrated in the hands of a few economic actors, who often offer their products through take-it-or-leave-it contracts. Data protection professionals will therefore need to evaluate elements such as what kinds of liability are excluded by their organization’s contract with a provider, what kinds of information are supplied, and whether their organization will be able to fully discharge its data protection duties under the terms of the contract. Those and other questions cannot be fully exhausted by a single training module, but the sessions of this module supply a starting point for finding out what aspects need to be verifying before hiring (or offering) an AI system or model in the market.

Session 6.2. Securing personal data for AI systems

By the end of this session, learners will be able to **distinguish** between various sources of personal data for AI systems and **examine** whether the organization has a legal basis for processing that data for the construction of an AI system.

Data is essential for AI systems and models. When we are talking about **machine learning** models, the rules that reside at the core of those models are derived from the statistical patterns present in their training data, which are then generalized. But even AI systems powered by other types of models, such as knowledge-based systems, will still need input data to **generate their outputs**, which often amount to data themselves. So, to the extent that those forms of data relate to identified or identified natural persons, an AI system or model will be steeped in personal data.

However, the data needed to create an AI system or model is not always easy to come by. This is especially true when it comes to large-scale technologies such as large language models, which have already been trained on basically every freely available piece of data available on the internet (Kuru 2024). But it is also the case for smaller models. For example, **InnovaHospital** might struggle to develop an AI-based predictor for a given illness if there are only a few known cases of that disease.

Additionally, not all data is made equal. Some sources might accurately capture an object of analysis, while others might supply badly measured or even deliberately misleading information. For example, data scraped from an online forum will likely reflect the biases and prejudices of the users of that forum. This is why some of the major players in AI technologies have emphasized the need for high-quality data as a competitive differential.

In this context, any organization wishing to develop an AI system or model—for its own use or for others—needs to consider how much data it has available for that purpose. It might be the case that an organization has enormous amounts of data it can apply to this new purpose. But it might also be the case that an organization must acquire new sources of data, either because it lacks the precise kind of information it needs or because existing sources are inadequate. In both cases, the organization will need to fulfil some legal requirements before it can use that data.

During the design and development stage of an AI system, personal data is most likely to be processed in the training processes of a machine learning model. As discussed in Unit 2 of this training module, many of the modern applications of AI rely on machine learning, and as such their decision rules are learned from data. So, if a model is

expected to take personal data as input or generate it as output, its training will likely require some personal data.

All legal bases for processing listed in Article 6(1) GDPR remain theoretically viable for AI systems and models. However, many of them stipulate that the processing must be "necessary" for the performance of some task. Given the narrow interpretation of necessity that prevails in data protection law, Articles 6(1)(b–e) GDPR are unlikely to sustain large-scale processing for the use of AI. In most cases, this means data controllers will need to rely either on the data subject's consent or in the presence of a legitimate interest that justify processing. Both options demand considerable work from the organization.

Consent as a legal basis for AI training

Two main difficulties emerge when one seeks consent in the context of training AI:

1. **Scale**: training a large-scale AI system might require data from thousands or even millions of people. The organization would need to identify them and contact them for securing consent.
2. **Complexity***:* once all parties are identified, consent needs to be freely given, specific, informed, and unambiguous. All those conditions can be problematic in the training of an AI system:
    a. The information and power between data subjects and the organizations that can train large AI systems can blur the lines of consent. For example, a student might authorize processing because they are afraid of being singled out by **UNw** in the future.
    b. It might be difficult to provide specific information about how AI is to be used in the training process, given the current limitations to our understanding of what goes on within an AI system.[7]

The collection of consent from data subjects must address those difficulties. Otherwise, consent might not be deemed valid for failing to meet one or more of the legal requirements in Article 7 GDPR. Additional requirements for consent might apply considering the sector in which the data is processed, such as the informed consent requirement for medical data.

Legitimate interest as a basis for training AI systems

As an alternative to the difficulties of consent, some organizations have considered the use of the **legitimate interest** basis[8] for training AI systems. This legal basis also authorizes processing that is "necessary" for a purpose: the pursuit of legitimate interests by the controllers or a third part. That basis does not apply when such interests

[7] See also Unit 11 of this training module.
[8] Article 6(1)(f) GDPR.

are overridden by the interests or fundamental rights of the data subject. For example, the pursuit of an economic interest might not justify severe intrusions into the right to a private life, especially the life of a children. Furthermore, it is unsuitable for the processing of special categories of personal data, as Article 9(2) GDPR does not feature a general clause on legitimate interest.[9]

In the absence of such an override, the controller is required to **weigh the legitimate interests being pursued against the rights and interests that might be affected by treatment** (Sartor and Lagioia 2020). This weighing follows the same procedure used for legitimate interest in other contexts. What changes is that it must consider AI-specific risks, such as the ones examined in Units 3 and 4 of this training module. As such, legitimate interest might allow more flexibility for AI developers, at the cost of requiring them to exercise more responsibility in analysing the consequences of their development (Kramcsák 2023). Future guidance from data protection authorities will likely clarify the use of this legal basis. In the meantime, data controllers need to have particular caution when relying on it for AI.

Legal bases for the reuse of personal data

Sometimes, an organization might want to use data it already has collected for purposes other than the construction of an AI system. Such a processing must, of course, have a legal basis in the GDPR. If that basis is not the consent of the data subject,[10] the data controller must observe **whether that new purpose is compatible with the purpose of original collection**.

Article 6(4) GDPR provides an open list of criteria that must be considered in this context, such as the link between the purposes of original collection and the intended further processing or the existence of appropriate safeguards. Those criteria must be assessed in AI training, just as they would be in any other processing.

At least three AI-specific factors must be taken into account in this assessment:

- When considering the link between purposes, one must consider **the purpose for which the system or model is being trained**. For example, using personal data related to an individual's credit behaviour to train a credit scoring system is, in principle, justifiable. Using that same data for training a general-purpose model, less so.
- When assessing the consequences of further processing for data subjects, it is necessary to consider two types of consequences:

[9] Narrower clauses authorizing processing in some cases are present in that provision. However, the "necessity" requirement must be considered when deciding on the extent of training that can be carried out.

[10] Or Union or Member State law, in the context of the restrictions to data protection authorized in Article 23 GDPR.

    - Those emerging from the **training process** itself, such as the risk of leaks of the stored personal data.
    - The **impact that the trained AI system might have** on the data subjects whose data was used to train on.
- When determining the safeguards for processing, one must consider **technical and organizational** measures that address the risks it creates, such as those identified in Units 3 and 4 of this training module. For example, an organization might consider pseudonymizing any personal data it cannot anonymize before training.

In short, the reuse of personal data must take seriously the risks created both by the training process itself and by the subsequent use of the trained system or module.

Processing special categories of personal data in high-risk AI systems

As discussed above, there is no general clause allowing training for legitimate interest when it comes to the use of special categories of personal data. This means, for example, that data about a natural person's health cannot be processed on the grounds of legitimate interest. This creates a challenge for some kinds of application, such as the medical diagnosis tools envisaged by **InnovaHospital** and its partners.

To some extent, this challenge is mitigated by the hypotheses listed in Article 9(2) GDPR. Coming back to the example above, processing that is necessary for the purposes of medical diagnosis is covered by Article 9(2)(h) GDPR. However, **the term "necessary" must be read narrowly**, as a broad reading would reduce considerably the level of protection offered by the provision. As a result, some scholars have pointed out that there was considerable uncertainty about whether additional data could be used to mitigate the risks of biases in an AI system.[11]

Article 10(5) AI Act is aimed precisely at this gap. It allows the processing of special categories of personal data in the training of high-risk AI systems if that processing is necessary for detecting and correcting biases. Whenever this exception is invoked, the following conditions must be met:

> *(a) the bias detection and correction cannot be effectively fulfilled by processing other data, including synthetic or anonymised data;*
>
> *(b) the special categories of personal data are subject to technical limitations on the re-use of the personal data, and state-of-the-art security and privacy-preserving measures, including pseudonymisation;*
>
> *(c) the special categories of personal data are subject to measures to ensure that the personal data processed are secured, protected, subject to suitable safeguards,*

[11] see, for an overview, van Bekkum and Borgesius 2023.

including strict controls and documentation of the access, to avoid misuse and ensure that only authorised persons have access to those personal data with appropriate confidentiality obligations;

(d) the special categories of personal data are not to be transmitted, transferred or otherwise accessed by other parties;

(e) the special categories of personal data are deleted once the bias has been corrected or the personal data has reached the end of its retention period, whichever comes first;

(f) the records of processing activities pursuant to Regulations (EU) 2016/679 and (EU) 2018/1725 and Directive (EU) 2016/680 include the reasons why the processing of special categories of personal data was strictly necessary to detect and correct biases, and why that objective could not be achieved by processing other data.

Hence, the AI Act extends the possibilities for using special categories of personal data during the training process but imposes considerable constraints in doing so.

Session 6.3. Processing data in AI development

By the end of this session, learners will be able to **discuss** how the data protection principles are affected by AI technology and **identify** AI-specific data protection rules.

Once an organization secures a legal basis for all the data it intends to use, it still has data protection obligations. After all, data protection law does not merely specify when data can be processed. It also lays down requirements for processing. As discussed in the introduction to this Unit, those requirements must be observed both when the means for processing are determined and when actual processing takes place. Hence, the design and the development of AI systems are highly relevant from a data protection perspective.

Briefly recapitulating the discussion in Unit 2 of this training module, it can be useful to distinguish between the various elements of an AI system that are determined at this life cycle stage:

- The **components** that will form an AI system, such as AI models, the hardware that will be used to execute those models, or its interfaces with other systems.
- The **AI model(s)** that will power that system.
- If the organization is programming its own model, or finetuning an existing one:
    - The **training data** from which the model will infer its rules.
    - The **learning process** it will follow for that inference.
- The **validation data** against which the model will be assessed.

- The **interfaces** through which potential users might use the system.

Those and other choices are directly relevant for data protection when they involve personal data during the training stage. Even if that is not the case, they might be relevant if the system is intended to process personal data. Either way, the involvement of a data protection professional at the design and development stage can prevent many headaches later.

In this session, we will cover issues that appear while interpreting the GDPR rules and principles in contexts involving AI. By necessity, any such treatment is partial, as many factors depend on the specifics of where AI will be developed and used, as well as on the techniques being used. Some examples will be added to mitigate this factor, suggesting how the learner can deepen the general guidelines offered here.

Applying data protection principles in design choices

The starting point for this inquiry is Article 5 GDPR, which lays down **general principles** applicable whenever personal data is processed. As principles, they do not offer clean-cut commands that one can either obey or not. Instead, their legal content is more abstract. They outline certain values that must be promoted, acknowledging that those values might be weighed differently in each case (Roßnagel and Richter 2023). For example, the new legal basis for processing created on Article 10(5) AI Act reflects the idea that, in the context of high-risk AI systems, fairness in the AI outputs can take precedence over strict data minimization. Compliance with data protection principles thus requires a balancing act between values in a concrete context.

What changes when AI comes into play? The general logic of principles remains the same, but AI systems and models transform the technical context of processing. Their impact can be felt in each of the GDPR's data protection principles.

Lawfulness, fairness, and transparency (Article 5(1)(a) GDPR)

The principle of lawfulness emerges as a cross-cutting principle. It establishes that any processing must both be allowed by law and follow the applicable legal requirements (Roßnagel and Richter 2023). Article 5(1)(a) GDPR highlights two facets of lawfulness: fairness and transparency. Both are affected by the use of AI.

When it comes to fairness, a developer must ensure two interrelated goals. It must ensure the fair processing of any data processed for training the AI system. That is, the developer must act in a way that justify the trust of the data subjects whose data is used in training. For example, if **InnovaHospital** decides to use patient data, it must do so in a way that does not mislead patients, provide adequate safeguards for their data, and does not harm them.

Developers must also ensure the fairness of the finished system or model. In particular, this principle compels developers to mitigate (or even eliminate) potential sources of algorithmic biases that might harm the rights and interests of those who will be affected by the use of an AI system. For example, fair processing in the context of **UNw**'s AI systems would require the university to adopt metrics to detect whether the system has a disparate impact on some group of students (for example, by discriminating against female students).[12] Because the construction of the AI system sets, to a large extent, the means of its future processing, the data protection principles must also be observed as technical choices are made.

The principle of transparency is analysed more closely in Unit 11 of this training module. For the time being, it suffices to say that it obliges developers to not just care about the transparency of how they process data but also about the transparency of further processing done with the AI system.

Purpose limitation (Article 5(1)(b) GDPR)

The principle of purpose limitation means that data must be collected for specified, explicit, and legitimate purposes, and that any future processing must not be incompatible with the original purpose. Its implications for the development process were unpacked in the previous session.

Data minimization (Article 5(1)(c) GDPR)

The data minimization principle establishes that personal data must be adequate, relevant, and limited to what is necessary for the purposes of processing. Each of these elements has implications for the use of AI technologies.

Regarding **adequateness***,* the developer must ensure they are using data of sufficient quality for the task at hand. For high-risk AI systems, this principle is further specified in Article 10 AI Act, and the measures presented therein (discussed below) can be a useful guide for organizations developing other kinds of systems or models.

The **relevance** element, in turn, suggests a developer should be able to tell whether and how the data they are bringing to the training process is relevant. For example, if **UNw** wants to predict the performance of its students in the courses they are taking, it has little reason to acquire training data from a broker that has collected information on the social media habits of those students.

Finally, the **necessity** element suggests that developers should not use a data-intensive solution when a solution that requires less data is available. In the context of automated decision-making, for example, it has been argued that complex black box models do not always perform better than simpler alternatives (see, e.g., Semenova et

[12] On metrics, see Unit 7 of this training module.

al. 2022). Whenever that is the case, the developer would do well to consider the advantages of using the more complex model. But, if simpler models cannot achieve the same result, this principle is not, in itself, an obstacle to the use of data-intensive AI.

Accuracy (Article 5(1)(d) GDPR)

We will examine this principle in more detail in Unit 7, where we discuss metrics that can be used to capture accuracy. Once again, this principle means that accuracy must be ensured both for the data used during the training process and for the AI system (or model) that will produce personal data in future uses.

Storage limitation (Article 5(1)(e) GDPR)

In an AI context, this principle relates mostly to the data surrounding the system or model itself. The training, test, and validation data are all subject to storage limitation, as well as the input and output data fed to the finalized system. If the AI model has some degree of memorization of personal data,[13] then the developer must also include mechanisms to ensure that the memorized data will not outlive its necessity.

Integrity and confidentiality (Article 5(1)(f) GDPR)

In an AI context, this principle entails that developers must attend to the security risks outlined in Unit 3 of this training module. For example, any organization developing an AI system must consider whether their system is vulnerable to model inversion attacks that would allow the extraction of personal data. If that is the case, mitigation measures become necessary.

Additional obligations for high-risk AI systems

The general principles outlined above are given concreteness in the GDPR's rules. In particular, Articles 25 and 32 GDPR require the developers of AI systems to adopt technical and organizational measures that implement those principles, as we will discuss in Unit 12 of this training module. Data subject rights, which are covered in Unit 8 of this training module, also are guided by those principles. Before wrapping up this unit, we will now briefly discuss the data management obligations introduced by the AI Act.

Under Article 10 AI Act, the provider of a high-risk AI system must adopt a variety of data governance measures. Article 10(2) AI Act defines a set of data governance and management practices that must be observed. Any provider using data in training a high-risk AI system must have oversight and control of how data is used, especially:

(a) the relevant design choices;

[13] See Session 2.1 of this training module.

(b) data collection processes and the origin of data, and in the case of personal data, the original purpose of the data collection;

(c) relevant data-preparation processing operations, such as annotation, labelling, cleaning, updating, enrichment and aggregation;

(d) the formulation of assumptions, in particular with respect to the information that the data are supposed to measure and represent;

(e) an assessment of the availability, quantity and suitability of the data sets that are needed;

(f) examination in view of possible biases that are likely to affect the health and safety of persons, have a negative impact on fundamental rights or lead to discrimination prohibited under Union law, especially where data outputs influence inputs for future operations;

(g) appropriate measures to detect, prevent and mitigate possible biases identified according to point (f);

(h) the identification of relevant data gaps or shortcomings that prevent compliance with [the AI Act], and how those gaps and shortcomings can be addressed.

Data quality requirements appear in Article 10(3) AI Act. Under this provision, the training, validation, and data sets must be:

- Relevant
- Sufficiently representative
- To the extent possible
    - Free of errors
    - Complete in view of the intended purpose

The relative character of the latter two obligations is crucial, given that perfect data does not exist. Nonetheless, this obligation forces providers of high-risk AI systems to pursue completeness and accuracy in their datasets.

Finally, Article 10(4) AI Act requires providers to use data sets that take into account some contextual elements. To the extent that those elements are required by the system's purpose, the datasets must consider characteristics or elements that are "particular to the specific geographical, contextual, behavioural or functional setting within which the high-risk AI system is intended to be used." For example, **UNw** must take into account the socioeconomical characteristics of its student body, while **InnovaHospital** must consider (among other things) whether some diseases it wants to diagnose with AI are affected by geographic factors.

Those measures are meant as quality criteria for the data used in the training process of an AI system. Being targeted at high-risk AI systems, they are not mandatory for any other type of system or model. Even so, they represent best practices that organizations might want to consider as a starting point for designing their own data governance architecture.

Conclusion to Unit 6

As of late 2024, the Irish Data Protection Commission has requested that the European Data Protection Board produce an opinion on the processing of personal data during AI development and training. The resulting opinion has been adopted by the EDPB on 17 December 2024, right as the first version of this training module was finalized. Therefore, the guidance offered above should be read in light of those new regulatory guidelines. Nonetheless, the discussions above offer a high-level overview of data protection issues that appear when training or developing an AI system.

The key takeaways from the previous discussion are:

1. Unless it is acting strictly under detailed instructions from a buyer, a developer will likely qualify as the data controller of the data it processes in the training process.
2. Depending on the circumstances under which an AI system or model is commercialized, the developer might also qualify as a joint controller for subsequent data processing.
3. Design choices must ensure the protection of personal data both regarding the processing that takes place in the training process and the future processing that will be done with an AI system or model.
4. Most uses of data during the training processes will likely be based on consent or legitimate interests, which means developers need to pay close attention to whether the requirements of those bases are satisfied.
5. The particularities of AI affect the interpretation of the various data protection principles, which nonetheless remain in force.
6. The data governance measures in Article 10 AI Act are obligatory for high-risk AI systems but they can also be useful for developers of other systems.

The three sessions of the Unit illustrate how data protection professionals can play a vital role in shaping the development process. If they collaborate closely with technical experts, they can do more than pointing out the unlawfulness of processing. They can help the organization find lawful bases for using the data it already has available, propose safeguards to ensure AI is used in a way that respects the rights of data subjects (including but not limited to the right to data protection) and make sure that design decisions are properly documented for future demonstrations of compliance.

Each of those practices contribute to the lawful use of the developed AI systems and models, be it by the developer itself or by third parties.

Prompt for reflection

Securing a valid legal basis for processing personal data is critical during AI development. However, consent and legitimate interest both present challenges, particularly for large-scale or high-risk systems.

- In your opinion, which legal basis (consent or legitimate interest) is more practical for training AI models in different sectors (e.g., education, healthcare, or commercial AI)? Why?
- Reflect on a case study like **DigiToys** or **InnovaHospital**—what factors should these organizations consider when choosing a legal basis?

References

Marvin van Bekkum and Frederik Zuiderveen Borgesius, 'Using Sensitive Data to Prevent Discrimination by Artificial Intelligence: Does the GDPR Need a New Exception?' (2023) 48 Comput Law Secur Rev 105770.

CNIL's Q&A on the Use of Generative AI (18 July 2024). Accessed 26 September 2024.

CNIL, Determining the legal qualification of AI system providers (07 June 2024).

Data Protection Commission. AI, Large Language Models and Data Protection (18 July 2024). Accessed 26 September 2024.

Ralf Kneuper, *Software Processes and Life Cycle Models: An Introduction to Modelling, Using and Managing Agile, Plan-Driven and Hybrid Processes* (Springer International Publishing 2018).

Pablo Trigo Kramcsák, 'Can Legitimate Interest Be an Appropriate Lawful Basis for Processing Artificial Intelligence Training Datasets?' (2023) 48 Computer Law & Security Review 105765.

David Lehr and Paul Ohm, 'Playing with the Data: What Legal Scholars Should Learn About Machine Learning' (2017) 51 UCDL Rev 653.

Silverio Martínez-Fernández and others, 'Software Engineering for AI-Based Systems: A Survey' (2022) 31 ACM Trans Softw Eng Methodol 1.

Vagelis Papakonstantinou and Paul de Hert, 'Art. 2. Material Scope' in Indra Spiecker gen. Döhmann and others (eds), *General Data Protection Regulation. Article-by-Article Commentary* (Beck; Nomos; Hart Publishing 2023).

Alexander Roßnagel and Philipp Richter, ‘Art. 5. Principles relating to processing of personal data’ in Indra Spiecker gen. Döhmann and others (eds), *General Data Protection Regulation: Article-by-article commentary* (Beck; Nomos; Hart Publishing 2023).

Giovanni Sartor and Francesca Lagioia, ‘The Impact of the General Data Protection Regulation (GDPR) on Artificial Intelligence’ (European Parliamentary Research Service, 2020).

Lesia Semenova and others, ‘On the Existence of Simpler Machine Learning Models’ in *2022 ACM Conference on Fairness, Accountability, and Transparency* (FAccT ’22, New York, NY, USA, Association for Computing Machinery 21 June 2022).

Suzanne Snoek and Isabel Barberá, ‘From Inception to Retirement: Addressing Bias Throughout the Lifecycle of AI Systems. A Practical Guide’ (Rhite and Radboud Universiteit 5 September 2024).

Rob van der Veer, ‘ISO/IEC 5338: Get to know the global standard on AI systems’ *Software Improvement Group*. Accessed 26 September 2024.

Unit 7. Verification and Validation of AI Systems and Models

By the end of this unit, learners will be able to:

1. **Distinguish** between various approaches to examining an AI system, such as software testing, evaluation of metrics, and audits.
2. **Identify** moments when assessment is needed, before and after the initial deployment of a system; and
3. **Incorporate** data protection questions into those assessments.

In a linear picture of the AI life cycle, the verification and validation stage takes place once the major software development activities take place. At this point, software developers (as well as specialized QA professionals) evaluate the mostly finished system (or model) to determine whether it is ready for use. They do so by subjecting it to a variety of tests, which are meant to evaluate whether the system or model meets the requirements identified in the inception stage.[1] If the system fails to meet those standards, it goes back to the design and development stage for adjustments.[2] Otherwise, it is deemed ready to be sold or put into service. In this Unit, we examine how those practices matter for data protection compliance.

Before that, we need to consider what it means for an AI system or model to be "ready." There is **no tried-and-true formula** that allows us to determine when a system has met the requirements that motivated its original design. Even if those requirements have been defined in objective terms, such as "the system must achieve 99.9999% accuracy according to [an established metric]," those might no longer be relevant by the time the software system is ready. Sometimes this happens because technology has evolved and what was previously acceptable is now a deficient performance. Sometimes the problem resides in the relevant criteria themselves, which are no longer relevant for the new context of an organization. Ultimately, what makes a software ready is the developer's decision to commercialize it (or put it into service).

That decision is, more often than not, influenced by **external factors** such as business needs or an attempt to catch up with the hype surrounding AI. Still, **the criteria laid down in the requirements stage**—as well as any subsequent updates—can play a part in the organization's decision-making process. Incorporating data protection considerations into those factors is therefore one way to increase their weigh in the AI development process.

A data protection professional can make a business case for that integration. First, addressing data protection risks at this stage can help organizations avoid issues once

[1] See Unit 5 above.
[2] See Unit 6 above.

a system has been deployed, thus **reducing the costs of compliance** with the GDPR's requirements for data protection and security by design.[3] Second, **the AI Act reinforces this general requirement** by stipulating conditions (including data protection requirements) that must be met before high-risk AI systems and general-purpose AI models with systemic risk can be placed on the market. Third, active compliance with data protection law can have **commercial advantages**, in particular by making an AI product more attractive to business clients who will themselves need to comply with data protection requirements. A developer would do well to integrate data protection considerations into all stages of its development cycle rather than dealing with problems as they emerge.

Accordingly, this unit discusses three moments within this life cycle stage in which data protection can be a relevant consideration. **Session 7.1** provides an overview of metrics track various properties related to data protection, such as accuracy, fairness, and data minimization. **Session 7.2** discusses tests and benchmarks that can be used to evaluate those metrics. Finally, **Session 7.3** discusses how audits can help evaluate systems before and after deployment.

Session 7.1. Measuring data protection

> By the end of this session, learners will be able to **exemplify** metrics that can be used to support compliance with data protection and *describe* their limits.

Performance measurement is part of the software development process. At various moments during that process, software developers can measure various indicators that describe aspects of the development process. Some of these indicators can be used to track functional requirements, such as the accuracy level of an AI model for a particular task. Others can be used to track non-functional requirements, such as the amount of data used for training the model or the amount of energy it consumes during the training process. In this session, we will discuss how those measurements can be applied to track data protection requirements.

Generally, there is **no obligation to track specific indicators** when it comes to AI systems. There is no mandatory threshold for indicators such as accuracy, either. This is because such measurements are highly contextual. One type of measurement that can be useful in one context might be unhelpful in another. For example, there is no sense in measuring the use of training data if one is using an expert system that is not trained on data.

[3] Articles 25 and 32 GDPR, respectively.

Likewise, thresholds that are perfectly acceptable in a particular context might be unacceptable elsewhere. If **InnovaHospital** creates a system that diagnoses a complex disease in 99.99% of the cases, this might be an improvement over the performance of human physicians. But if a large social network creates an automated content moderation system with the same level of accuracy, it will result in thousands, maybe even millions of valid posts being removed by the system.

There is **no indicator or set of indicators that is guaranteed to be useful in all cases**. Instead, an organization must look at the risks potentially created (or amplified) by their AI system or model and choose what indicators can be relevant for their problem.

Usually, this means organizations will need to rely on a broad range of indicators, each capturing a different aspect of the AI system or model. Even considered in aggregate, those indicators will only **offer a partial view of their object**:

- Some relevant aspects of the impact of an AI system might not be amenable to metrification. For example, one might argue that core aspects of human personality cannot be quantified (Hildebrandt 2019).
- Alternatively, something might be measurable in theory, but not measured in practice. This can happen, for instance, if somebody decides not to measure a certain indicator, or if measurement is too expensive or otherwise unfeasible.
- An indicator might be inadequate for the task at hand. This is likely to be the case when a system is faced with a scenario that is far away from its usual range of operation. For example, during the Chernobyl disaster, the radiation counters available to first responders could only ascertain the radiation levels were above 3.6 Roentgen per hour, which was the limit of their instruments, but actual levels were much higher.
- An accurately measured indicator is of no help if no one bothers to read it.

Measurement is not enough to ensure compliance with data protection requirements. But a proper use of a diverse set of quantitative and qualitative metrics can help organizations identify risks associated with their AI system or model, either before development or after deployment. Hence, the use of data protection metrics can be a powerful tool for compliance.

When it comes to high-risk AI systems, Article 15(1) AI Act mandates that such systems must have “appropriate” levels of accuracy, robustness, and cybersecurity. It does not define what counts as appropriate; as discussed above, what is adequate in a context might be awful in another. Instead, Article 15(2) AI Act stipulates that the Commission, in cooperation with other stakeholders, shall encourage the development of benchmarks and measurement methodologies. Likewise, sector-specific rules and industry standards will provide more information about acceptable thresholds in particular

contexts.[4] Once those definitions become available, developers of high-risk AI systems must ensure the relevant levels of accuracy, robustness, and cybersecurity. For developers of other AI systems and models, the applicable rule might not be mandatory, but it can still offer guidance for determining what levels are appropriate for their application.

Furthermore, Article 15(1) AI Act also requires that high-risk AI systems be consistent in those respects throughout the entire life cycle. That is, they must not suffer substantial degradation when it comes to those properties. To ensure that is the case, developers will need to track their AI systems and models after deployment, potentially rolling out updates if changes in technology or context make things worse. Compliance with this requirement must consider the data protection factors discussed above whenever the high-risk AI system involves the processing of personal data.

As of the end of 2024, there is limited agreement on what metrics and indicators are suitable for tracking various aspects of AI systems and models. In Unit 14 of this training module, we discuss instruments that are likely to provide more clarity in this regard, such as harmonized technical standards and codes of practice sponsored by the European Commission. In the meantime, it will be useful to define certain metrics and indicators that can support data protection assessments.

Measuring accuracy

The term “accuracy” often appears in the context of AI technologies. It is used, for example, in Article 15 AI Act, which oblige the providers of high-risk AI systems to ensure that their systems are sufficiently accurate for their purposes. In the broadest sense, this requirement for accuracy can be understood as a requirement that the AI system performs as close as possible to the results one would expect in that context. To measure that, technical experts have proposed a variety of indicators.

Classification accuracy

Some of those indicators are tailored for **classification** tasks. A classification task is a scenario in which an AI system is expected to assign an output to one of two (or more) possible classes. For example, an image recognition system might distinguish between photos that feature a dog and photos without a dog. When the AI system’s goal is formulated like that, its performance can be measured through some specific measures.

Those measures are often built from the same building blocks, that is, they offer diverse ways to combine certain indicators. For binary classification problems (in which an object can belong to one of two classes), a few indicators are common:

[4] See Unit 14 of this training module.

- **Precision** is the likelihood that an object assigned to a class actually belongs to that class. For example, if the **UNw** university builds a classifier for predicting student dropout rates, its precision can be measured by computing how many of the predicted dropouts dropped out.
- **Recall**, also known as **sensitivity**, refers to the likelihood that the system will correctly label the elements belonging to a given class. In the previous example, for instance, recall would refer to how many actual dropouts were identified.

An example of an indicator built from those two indicators is the F1 score that is commonly used in binary classification problems. That score is calculated as the harmonic mean of a system's precision and recall.

Accuracy in regression

Not all problems solved by AI are classification problems. Some applications, for instance, focus on what is usually called **regression**. That is, an AI system is expected to predict a future value of a variable based on its present value. For example, **DigiToys** might use a regression tool to forecast its future sales based on data about its current performance and other relevant market values.

In a regression problem, it is very unlikely that an AI system will predict the exact value of the target variable. This does not mean that all errors are all the same. If **DigiToys**'s predictor gets the revenue forecast wrong by some million Euros, the company is likely to have serious problems. If it gets things wrong by a few cents, the impact is much less relevant. As such, indicators for evaluating regression performance need to account for the distance between the expected result and the real result.

One common indicator is the **mean average error (MAE)**. This indicator is relatively simple to calculate. One calculates the difference between the expected value and the value that was observed in each case, takes the absolute value of that difference (that is, ignores the sign), and then gets the mean between all those values. This metric's simplicity is an advantage for calculation, and it can be more easily explained. However, it treats all errors equally, which might not be desirable in all circumstances. For example, one cannot easily distinguish between a scenario where a high MAE is the result of high errors in all cases or of a single outlier that is incredibly wrong.

To compensate for this shortcoming, practitioners often rely on other metrics. One such metric is the **mean squared error (MSE)**. The MSE is calculated like the MAE, with one difference: before computing the mean, one takes the square of all the differences. By doing so, one ensures that large errors will become even larger, while smaller errors become vanishingly small. Thus, reducing MSE would show that a system is less prone to significant deviation.

Robustness metrics

The robustness of an AI system or model refers to its capability to continue operating as expected even when it faces errors and unexpected inputs. This means a robust system will continue to be reliable under varying conditions. Software engineering professionals have developed a variety of indicators for a system's robustness, many of which are related to time:

- The **mean time between failures** (MTBF) counts how much time a system can operate without undergoing an incident that disrupts its operation.
- The **recovery time**, instead, focuses on how fast a system can recover from such disruption.

One might, for example, want to track the types of failures to which a system is exposed. Different AI systems or models might fail in diverse ways, and the impact of each kind of failure also varies depending on the context of use:

- A system used for medical diagnoses at **InnovaHospital** faces high stakes, as it must remain functional when faced with sudden data influxes or when exposed to data that is not particularly accurate, as the conditions for measurement are not always ideal in practice.
- The systems produced by **DigiToys** must be able to cope with the unpredictability of child behaviour. For instance, a learning puzzle must withstand incorrect or inconsistent input without freezing or providing nonsensical feedback.
- An AI system operated by **UNw** might need to deal with huge variations in its operation volume. For example, the demand for tutor chatbots is likely to grow considerably right before the university's exams.

In those cases, robustness could be measured by context-specific quantities, such as indicators that capture how much the system's operation is impacted by minor changes in the input data. Those might be complemented by context-specific qualitative indicators, such as those derived from customer satisfaction evaluations.

Cybersecurity metrics

Over the past decades, cybersecurity professionals have developed a variety of specialized metrics to capture several aspects of security. It would not be feasible to cover them all in detail, but the introduction to cybersecurity in Unit 3 of this training module already suggests a few measurable aspects.

In terms of quantitative measurements, one can look at the main objects of cybersecurity. It might be possible to measure the number of identified vulnerabilities, or the time it takes to patch a vulnerability once a fix is available. Measurements might also cover the organization's cybersecurity practices, for example by capturing the

frequency with which the organization searches for vulnerabilities, or the time it takes to respond to incidents or carry out adversarial tests of its systems.

Other measurements might not be so easy to translate into numbers but are nonetheless relevant. An organization might evaluate the suitability of its technical measures (such as encryption) and the extent to which it complies to existing cybersecurity standards. By defining qualitative and quantitative targets beforehand, an organization can gain a more holistic perspective on its security situation.

Session 7.2. Evaluating AI software for data protection issues

> By the end of this session, learners will be able to **describe** different approaches for software testing and **identify** when they are legally required for AI systems.

Software metrics, such as those discussed in the previous session, can be used to describe an AI system or model's operation and evaluate how it changes over time. As such, they are particularly useful for tracking its post-deployment life cycle. However, measurements are also important *before* a system is cleared for deployment. On the one hand, measuring properties of a system or model before deployment might tell us that the system requires further development before it is ready for use. On the other hand, those initial measurements offer a baseline against which one can compare future changes in the AI system. To obtain those initial values for the relevant indicators, a developer can follow software testing practices.

In EU data protection law, software testing is required under Article 32(1)(d) GDPR, which requires "testing, assessing and evaluating" of technical and organizational measures for secure processing. Article 25(1) GDPR, on data protection by design, does not feature an explicit mention to software testing. However, this provision requires data controllers to address risks to data protection principles that can emerge from processing. It is difficult to see how such risks can be identified without comprehensive testing.

Acknowledging that, the AI Act provides explicit testing requirements for high-risk AI systems and general-purpose AI models with systemic risk. For high-risk systems, Article 17(1)(d) AI Act obliges providers to define procedures for examining, testing, and validating the system throughout the entire life cycle. For general-purpose AI models, Article 55(1)(a) AI Act obliges providers to perform model evaluation "in accordance with standardised protocols and tools reflecting the state of the art". Those two provisions add more details to the general testing requirement that can be read in the GDPR.

In this session, we will discuss how those tests can be carried out.

Levels of software testing

Software engineers have developed various approaches for systematically testing computer programs. Those tests can be used to evaluate various aspects of a system, capturing information about (for instance) the metrics we discussed in the previous session. A comprehensive testing suite might therefore ensure that an AI solution is functional, reliable, and compatible with other software and hardware components.

One can distinguish between four types of tests:

- **Unit tests** focus on verifying the functionality of individual components in isolation. For example, **DigiToys** might evaluate the speech recognition unit in an AI-powered doll, ensuring it correctly identifies a single spoken command in controlled conditions.
- **System tests** assess the AI system as a whole, ensuring that all components work together as intended in a realistic environment. As an example, **UNw** might perform system tests on an AI scheduling tool by simulating real-world use cases, such as assigning classrooms and faculty to hundreds of overlapping courses during peak enrolment periods.
- **Integration tests** focus on ensuring compatibility and proper communication between different components or systems. At **InnovaHospital**, for example, integration tests might confirm that a diagnostic AI system retrieves real-time patient data from hospital servers without introducing delays or errors.
- **Acceptance tests** are conducted to determine whether the AI system meets the user's requirements and is ready for deployment. These tests typically involve end users interacting with the system in a simulated or real environment. For instance, **DigiToys** could have parents and children test an AI educational toy to assess whether its interaction is engaging, safe, and aligned with educational goals.

Those tests deal with distinct aspects of an AI system, and as such they complement one another. By combining them, organizations can ensure that AI systems and models not only function correctly but also meet real-world expectations and requirements. However, implementing those levels of testing in concrete scenarios will likely require the use of techniques that attend to the specifics of AI technologies.[5]

Software benchmarking and its use for AI

Another way to evaluate computer systems is to subject them to pre-defined benchmarks. In the context of AI, a **benchmark** typically involves a dataset, a set of tasks, or performance metrics that an AI system is tested against. Benchmarks provide

[5] For more details on technical measures, see, among others, Enrico Glerean, *Elements of Secure AI Systems.*

a standardized way to measure how well an AI system performs specific tasks, allowing developers to identify strengths, weaknesses, and areas for improvement.

This approach has been embraced by the AI Act, which establishes that the classification of a general-purpose AI model as a general-purpose AI model with systemic risk depends on whether the model meets established benchmarks to that end.[6] However, the utility of benchmarks does not end with this classification: at least in theory, benchmarks can be designed to evaluate several aspects of an AI system or model.

One example of benchmark available to AI developers is the MLPerf Training benchmark suite. This suite is formed by a variety of datasets and tasks, and it is meant to evaluate the time that a high-performance computer system takes to train an AI system that can reach a pre-defined level of quality at that task. The components of this benchmark suite are themselves benchmarks for specific problems. For example, ImageNet is a large dataset of labelled images that are used for evaluating the performance of image classifiers.

While benchmarks are invaluable for assessing and comparing AI systems, they are not without limitations. A key challenge is that benchmarks often measure performance in **controlled, idealized conditions** that may not reflect the complexities of real-world scenarios. For instance, an AI system trained and tested on the ImageNet dataset might perform well in the benchmark but fail to generalize to new, diverse images encountered in practice. This limitation is especially critical in high-stakes applications, such as healthcare or autonomous driving, where systems must operate reliably in unpredictable and dynamic environments.

Another limitation is that **benchmarks can oversimplify tasks**, focusing on narrow performance metrics that may not capture the full range of an AI system's capabilities or ethical implications. For example, accuracy metrics used in benchmarks often ignore fairness, robustness, or interpretability—factors that are crucial in domains like hiring or law enforcement. This narrow focus may inadvertently encourage developers to optimize for benchmark performance at the expense of these broader considerations.

Additionally, the useful of certain benchmarks might be eroded by some factors. As technology evolves, a specific benchmark might no longer be a stress test of a system’s capabilities, and thus become irrelevant. Another path to irrelevance is that sometimes an AI system might be trained on the benchmark’s dataset, or a dataset terribly similar to it. Doing so might ensure an exceedingly high performance on the benchmark that does not mean necessarily that the system is useful for real-world tasks. Organizations

[6] Article 51 AI Act.

can still benefit from adequate benchmarking, but they cannot afford to take results at face value.

Session 7.3. AI auditing requirements

By the end of this session, learners will be able to **distinguish** between black-box and white-box audits and **examine** what kind is suitable in each context.

So far, we have considered situations in which a software is tested by the organization that develops it. Such tests are an essential part of the development process. They are also desirable from a legal perspective, as they allow organizations to understand the risks that their AI systems or models might create, and thus anticipate legal exposure. Yet even a scrupulous internal process of testing might not capture all potential issues. Consequently, organizations developing software systems often rely on external audits of their products.

Audits can also be a useful tool for the governance of AI systems and models. But, given that the state of the art in AI technologies has evolved quickly over the past few years, techniques for auditing AI technologies are still relatively undeveloped as of the end of 2024. This situation is likely to change in the next few years, as considerable research is taking place about how to best audit technologies. For the time being, this session will focus on explaining fundamental concepts rather than presenting individual techniques that might soon become outdated.

The development of AI auditing techniques will be shaped, at least in part, by the legal requirements for audits. Many such requirements were already in the GDPR:

- Article 28 GDPR requires data processor to collaborate with audits conducted by (or on behalf of) the controller.
- Article 39(1)(b) GDPR mentions audits as part of the data protection officer's toolkit for monitoring compliance.
- Article 47(2)(j) GDPR refers to the need for data protection audits within groups of undertakings or enterprises engaged in a joint economic activity, to verify compliance with binding corporate rules on data protection.
- Article 58(1)(b) GDPR empowers data protection authorities to carry out investigations in the form of data protection audits.

To the extent that AI systems or models process personal data, be it during their training process or after deployment, they are covered by those audit powers.

Further audit requirements emerge from the AI Act's rules on high-risk AI systems. Under Article 74 AI Act, the market surveillance authorities are empowered to request data and documentation from the providers of high-risk AI systems, which they can use

for auditing purposes. Additionally, Annex VII AI Act details that certification bodies responsible for the third-party certification of some high-risk AI systems[7] must carry out periodic audits of the systems they certify. Those audits might require elements beyond those demanded by data protection law, but, to the extent that personal data is relevant to the system or model, they will need to include a data protection audit.

A recent paper by Casper, Ezell, and others (2024) distinguishes between three types of audits. In **white-box audits**, the auditor has access to the inner workings of an AI system or model, being able to change internal parameters and observe the consequences of that change. **Black-box audits** take place when an auditor has no access to the inner workings of an AI system or models but can provide inputs to that system or model and see which outputs it produces. Finally, **outside-the-box audits** analyse the development process and associated artefacts.

White-box audits as an ideal standard

A white-box audit, at least in theory, allows for the greatest level of scrutiny of an AI system or model. In this kind of audit, an auditor can thoroughly inspect the technical object in question. They have full visibility of the system (or model)'s internal parameters and can change them to see what happens with the system. This allows an auditor, for example, to detect whether the examples being tested have not been cherry-picked to show the system (or model) at its best performance, or to analyse how sensible that system (or model) is to external perturbations. A good white-box audit would therefore detect issues that would escape less intrusive means of observation.

In an ideal world, this would mean that AI systems and models are subject to white-box audits before and after deployment. There are, however, many obstacles to this approach in practice. From a practical standpoint, a comprehensive white-box audit is likely to take a long time, as AI systems and models have immense numbers of parameters that can be tinkered with. Given the technical expertise needed to make sense of the technical arrangements of even the simplest AI models, the **costs associated with such audits are likely to be high**.

Technical factors can also reduce the appeal of white-box audits in practice. Because AI systems are relatively recent, there is limited knowledge of what those audits should cover. To mitigate this factor, the European Data Protection Board has commissioned an AI Auditing project, which offers freely-accessible criteria that must be evaluated in an audit. Those factors can help an organization in setting up its audit requirements, which can be updated to match new technological developments.

White-box audits are further complicated by the technical arrangements of AI systems. Even if an organization allows an auditor to access every parameter it controls, **some**

[7] On certification, see Session 14.2 of this training module.

parts of an AI system might remain opaque to the audit. For example, if **InnovaHospital** uses ChatGPT to power a medical chatbot, an audit of that chatbot will not be able to access the inner workings of the large language model. It will still be able to access everything that the hospital has done *with* ChatGPT, but an important part of the AI system will be out of reach. So, the white-box audit in this case will need to deal with an unremovable black box.

Secrecy considerations might also reduce the attractiveness of a white-box audit. For example, the **DigiToys** company might fear that an external audit will result in a leak of its commercial strategy to competitors. This kind of risk can be mitigated by legal requirements of secrecy, such as contractual obligations concerning a company's trade secrets. However, an organization might be precluded from disclosing some information it holds due to agreements with third-party suppliers, too. A confidentiality clause in the contract with an AI provider, for instance, might prevent an organization from seeking a white-box audit.

Furthermore, white-box audits **are not mandated by law**. The transparency requirements in the GDPR and the AI Act do not go as far as to mandate disclosure of the AI system (or model)'s inner workings to external auditors.[8] A data processor might be obliged to undergo a white-box audit if that is stipulated in its contract with a data controller, and likewise, a joint controllership agreement can feature a requirement for this kind of audit. But, in the absence of such a contractual agreement or of a regulatory requirement, organizations can exercise their discretion on whether to pursue a white-box audit.

Black-box audits of AI systems

At first glance, a black-box audit might appear to be a suitable alternative to a white-box approach. In this kind of audit, an auditor inspects a system (or model) without having access to its inner workings. They can only **observe the system (or model)'s behaviours**: what outputs it generates for each input it receives. This is the approach followed by most techniques currently proposed for AI auditing, which try to exhaustively test the system without changing how it works. By doing so, those techniques would offer some guarantees about system behaviour while avoiding many of the pitfalls discussed above.

However, black-box audits can be inadequate for many real-world contexts. This is because they are vulnerable to several forms of **(deliberate or accidental) distortion**:

1. A black-box audit cannot assert that the system being tested is configured just like the system that will be examined in the real world.

[8] Even if regulatory authorities might have access to them by using their regulatory powers.

2. A black-box approach cannot, by definition, be as exhaustive as a white-box approach, as it cannot evaluate how the system's behaviour change when internal parameters are altered.
3. A black-box approach creates obstacles when it comes to finding the *source* of any issues detected during the inspection, as one cannot trace those issues to specific aspects of the system.

As such, a black-box audit can be a useful technique, but it might not cover all potential sources of legal issues with an AI system or model. A data protection professional will need to evaluate whether the guarantees offered by this kind of audit are enough in a particular context, considering the trade-off between clarity and feasibility. As a rule of thumb, the higher the risk associated with an application, the more access will be needed for an audit. Otherwise, an organization might miss valuable information and find itself with an unwarranted sense of security.

Outside-the-box audits

Unlike the previous kinds of audits, outside-the-box audits do not look at the AI system (or model) directly. Instead, they engage with **artefacts that are related to the technical object** they inspect. They look, in particular, to the documentation created during the development process of that AI system. For audits that take place later in a system's life cycle, the inspection might also cover documents about its deployment procedures. The idea is that those sources will contain information about the system itself.

This is the approach followed by the AI Act. Under its Article 43, some high-risk AI systems must undergo a third-party conformity assessment before they can be placed on the EU market. Such an assessment, as detailed in Annex VII AI Act, covers the documented process for quality management, as well as the system's technical documentation. No inspection of the system itself is carried out at this point.

Such an approach is prone to some of the issues with black-box audits. There are several reasons why a system's documentation might not match the actual system. If a system undergoes self-learning, its parameters will soon diverge from whatever is documented at a given moment. Even without that, it might be the case that some features of a system have not been entirely documented. This is more likely to be true in systems developed in accordance with agile processes, in which documentation is seen as less relevant than functionality. And, as the life cycle of a software system goes on, the documentation might become outdated considering software updates. The result is that software documentation tends to provide an incomplete picture of what happens within a system.

To mitigate those differences, the AI Act creates an obligation for providers of high-risk AI systems to keep updated the documents they are obliged to draft. But, even in the

absence of such an obligation, an organization would do well to keep its documents updated. After all, up-to-date documentation can be used as an element to demonstrate compliance with data protection requirements. To the extent that an organization takes care of software-related documents, an outside-the-box audit might provide useful insights about the AI system or model. Or at least some points that warrant further investigation by white-box (or black-box, if applicable) audits.

Conclusion to Unit 7

Verification and validation are continuous processes for AI technologies. A sensible provider will address its risks by carrying out tests and audits during the development of an AI system or model, while deployers would do well to extensively examine the systems they want to use before deployment. However, the same techniques described above can also be used for evaluating AI systems and models after the initial deployment. Such evaluations are, in fact, necessary, given both the possibility of technical changes to an AI system and the likely changes to the environment in which operates.

Based on the discussions in this unit, a data protection professional can carry out several types of intervention at this stage of the AI life cycle. They can:

- **Ensure** an organization selects a good mix of qualitative and quantitative metrics, which should cover various aspects of data protection compliance (such as accuracy, fairness, robustness, and cybersecurity).
- **Urge** organizations to keep track of those metrics throughout the life cycle, relying on tools such as up-to-date dashboards that concentrate information.
- **Participate** in the design of software testing protocols to ensure that factors relevant to data protection are covered by the tests.
- **Carry out** internal audits with a view to diagnosing data protection issues.
    - White-box audits offer a technical gold standard, but they might not always be feasible in practice.
    - Black-box audits are vulnerable to several limitations and possibilities of manipulation, but they might represent what is technically feasible in a certain context.
    - If black-box audits must be carried out, they need to be supplemented by outside-the-box methods, such as close analyses of software documentation.

By preparing robust practices for verification and validation, data protection professionals will help organizations comply with their legal duties throughout the entire life cycle.

Prompt for reflection

Discuss the advantages and limitations of black-box, white-box, and outside-the-box audits in ensuring compliance with data protection laws for high-risk AI systems. How would you approach auditing in cases where confidentiality agreements or technical opacity limit access to internal system parameters? Use examples from the case studies to ground your discussion.

References

Paul Ammann and Jeff Offutt, *Introduction to Software Testing* (2nd edn, Cambridge University Press 2016).

Stephen Casper and others, 'Black-Box Access is Insufficient for Rigorous AI Audits' in (ACM 6 March 2024) FAccT '24 2254.

Gemma Galdon Clavell, 'AI Auditing' (EDPB Supporting Pool of Experts, 2023).

David Lehr and Paul Ohm, 'Playing with the Data: What Legal Scholars Should Learn About Machine Learning' (2017) 51 UCDL Rev 653.

Mireille Hildebrandt, 'Privacy as Protection of the Incomputable Self: From Agnostic to Agonistic Machine Learning' (2019) 20 Theoretical Inquiries in Law 83.

Jakob Mökander and others, 'Conformity Assessments and Post-Market Monitoring: A Guide to the Role of Auditing in the Proposed European AI Regulation' (2022) 32 Minds & Machines 241.

Rob van der Veer, 'ISO/IEC 5338: Get to know the global standard on AI systems' *Software Improvement Group*. Accessed 26 September 2024.

Sandra Wachter and others, 'Why Fairness Cannot Be Automated: Bridging the Gap between EU Non-Discrimination Law and AI' (2021) 41 Computer Law & Security Review 105567.

Unit 8. The Deployment of an AI System

By the end of this unit, learners will be able to **devise** organizational procedures to mitigate risks to the protection of personal data that were not eliminated during software design.

The deployment stage is when an AI system is prepared for use. As such, it makes no sense to speak of the deployment of an AI model, given that we defined (in Unit 3) that a model is a component rather than a stand-alone product or service. At this point, the system's technical configurations have been mostly defined, except for some tasks that must be done at the moment a system is prepared for use, such as defining parameters. What remains is the work of preparing an organization for using the system: defining who will operate the AI system, how its outputs will be used, and so on. Those **organizational measures** set up the context in which the AI system is expected to affect a physical or virtual environment.

Data protection law is well aware of the impact that such organizational measures can have on the rights of data subjects. It empowers those subjects with rights they can oppose to specific instances of data processing (as we will discuss in this unit), and in doing so it creates obligations for data controllers. Those controllers are also obliged to adopt organizational measures—and not just technical ones—to address risks to data protection principles.[1]

For most AI systems, the governance of organizational measures is a matter of data protection law.[2] The AI Act, in line with its product safety pedigree, focuses on technical standards for AI systems and models. Still, Article 26 AI Act establishes some obligations for the deployers of high-risk AI systems, which must be implemented in a way that aligns with the general organizational duties imposed by the GDPR.

In this Unit, we will discuss three kinds of organizational duties related to data protection. **Session 8.1** discusses the AI literacy duty imposed by Article 4 AI Act, clarifying that its implementation can be a valuable tool for data protection compliance. **Session 8.2** discusses the challenges that the use of AI technologies creates to certain data subject rights, as well as possibilities of compensating technical obstacles with organizational measures. Finally, **Session 8.3** introduces measures that are meant to support the deployment of trustworthy AI by organization, such as regulatory sandboxes.

[1] As discussed in Unit 12 of the training module.
[2] And sector-specific laws, when applicable.

Session 8.1. The AI literacy obligation as an organizational measure

By the end of this session, learners will be able to **map** the various obligations following from the AI Act's AI literacy obligation and **exemplify** measures to foster it.

Of all the obligations created by the AI Act, only one applies to all AI systems: **AI literacy**. As expressed in Article 4 AI Act, this obligation means that:

> *Providers and deployers of AI systems shall take measures to ensure, to their best extent, a sufficient level of AI literacy of their staff and other persons dealing with the operation and use of AI systems on their behalf, taking into account their technical knowledge, experience, education and training and the context the AI systems are to be used in, and considering the persons or groups of persons on whom the AI systems are to be used.*

From the definition above, we can extract two elements to inform legal compliance. The first one is that Article 4 AI Act establishes an **obligation of best effort**. Providers and deployers of AI systems must show that they are adopting measures to foster literacy. They are not bound to any particular level of literacy, unless such a level follows from sector-specific legislation.

Instead, the target level of literacy is **contextual**. Literacy must be "sufficient" for the "operation and use" of an AI system. However, sufficiency is a broad concept: would it be enough just to teach operators what kinds of data they need to input in the system and what to do with the outputs? Or is it necessary to provide a deeper understanding of how the technologies work? The AI Act does not offer a clear-cut answer to this question, but the definition above indicates a series of factors that a provider or deployer must assess when designing their literacy measures.

Despite this ethereal formulation, one must keep in mind that Article 4 AI Act establishes an actual obligation. While the AI Act features some provisions that stimulate voluntary compliance,[3] this is not one of them, given the presence of the *shall* clause. It is true that Article 99 AI Act does not establish fines for breaches of the

[3] For example, Article 96 AI Act includes mechanisms that encourage providers of non-high-risk AI systems to voluntarily comply with some of the provisions for high-risk systems.

literacy duty. But it does not exclude the possibility of non-monetary sanctions.[4] Furthermore, literacy promotion is a measure that mitigates risk, and thus can lead to a reduced fine under Article 99(7) AI Act. The AI Act's literacy obligation is not entirely toothless.

Whenever an AI system operates on personal data, or is trained on it, that obligation must be read in line with existing data protection obligations. Articles 25 and 32 GDPR both require data controllers to adopt, among other things, organizational measures that are meant to address risks created by processing. Organizational measures refer to institutional processes regarding how data is processed, such as the definition of *who* can access data and *how* that data can be processed. As such, the proper implementation of those measures will require that persons within the organization have access to information about the system (and potentially, about its components). The AI literacy duty can therefore support compliance with data protection law.

The contents of AI literacy

The notion of "AI literacy" is formally defined in Article 3(56) AI Act. It encompasses the "skills, knowledge, and understanding" that allows providers, deployers, and affected persons to make informed decisions about AI and be aware of opportunities, risks, and potential harms from its use. This definition, interpreted in the light of data protection law, must guide compliance with Article 4 AI Act.

Compliance with the AI literacy requirement entails, in the first place, determining its contents. Given the formulation of "skills, knowledge, and understanding" above, it is **not enough to provide information about the existence of an AI system** and how a user can operate it. This kind of information is necessary (otherwise, people would not be able to discharge their duties or exercise their rights) but it must be accompanied by more **general, transferrable knowledge about what AI is and what it can (and cannot) do**.

The specific skills, knowledge, and understanding will depend on how a person is affected by the use of an AI system:

- A software developer involved in the process of deploying an AI system within an organization will need to understand some details about how that system operates, in order to diagnose errors and maintain the system.
- The persons operating that AI system, in turn, do not need to have a command of the technicalities of the system. But they need clarity about other aspects of the AI system, such as what role it plays within a context, what are its safety margins and known risks, and how to operate it correctly.

[4] Which are left to Member State legislation under Article 99(1) AI Act.

- Complaints channels within an organization need to be able to respond to requests from affected persons, such as the exercise of the data subject rights examined in Sessions 8.2 and 8.3 below.
    - The **data protection officer** is likely to be the first point of contact for this kind of request, especially when they are directly grounded on GDPR rights.
    - Other contact points in an organization, such as customer service or an ombuds, might also be contacted with complaints. In that case, they will need to liaise with DPOs to sort out data protection issues.
    - For that purpose, they will need **access to information** about AI: whether AI systems are or not used in each context, the decision logic of the systems in use, and about the data they use.

A data protection professional will need to have a clear view of AI systems and those operating them in order to design an adequate literacy programme. For that purpose, the AI inventory discussed in Session 5.2 of this training module might be particularly useful.

Once the targets of a literacy programme have been identified, they will likely need **tailored information programmes**. Given the several types of informational needs mapped above, different actors will need distinct levels of technical details and organizational context. For example, software developers can deal with more technical detail, but they are less likely to be familiar with the operational context in which the AI system will be used. A program that focuses on their knowledge is likely to require too much specialized knowledge to be useful for a non-technologist who simply uses the AI system. At the same time, it might lack contextual information that this person will need to do their job. Hence, there is no one-size-fits-all solution to AI literacy.

To that effect, Article 4 AI Act lists several factors that must be considered when determining the contents of an AI literacy program. The first set of factors refers to the system itself. The contents of a literacy programme must allow the target individuals to have a better understanding of the technologies used to power the AI systems being provided or deployed by the organization. For this purpose, an organization will likely need to provide some baseline knowledge about AI in general, but it can and should focus on the technologies it currently uses (or plans to use).

The second set of relevant factors in designing the literacy programme is that relating to the persons affected by the literacy programme. Under Article 4 AI Act, providers and deployers must foster literacy among the people they employ in AI-related roles and to external persons carrying out AI-related tasks on their behalf. For technical roles, this will likely mean a deep dive into the specific models and interfaces used for the system. For less technical roles, this means an explanation that is tailored to laypersons, who

need not discuss details but still require an overview of what is going on. In both cases, the focus is on providing a clear view of the opportunities and risks created by AI, in a way that allows people to make decisions that are compatible with their roles in the organization.

Finally, Article 4 AI Act requires that AI literacy considers the opportunities and risks that AI creates for the persons affected by the use of the AI system. This does not mean that an organization must foster literacy among those persons. Instead, it obliges the organization to teach people how to take the rights and interests of those persons into account. And, when one speaks of “affected persons,” it is likely that the AI system is processing personal data, which means the general requirements for the design of organizational measures laid down in the GDPR remain in force.[5]

One cannot foster AI literacy without creating literacy about how personal data is processed by and through AI systems within an organization. Therefore, data protection professionals can seize the AI literacy obligation as an opportunity to create awareness about the obligations of the various actors involved in the AI life cycle.

Promoting AI literacy within an organization

Once an organization has a clear view of who is involved in the deployment and development of AI systems, and of the information that is relevant for the tasks of those actors, it can start designing literacy programmes that meet their needs. There is no established formula for AI literacy, yet, but one can already prescribe useful steps for programme design.

First, one needs to know the **starting point** for the literacy programme. Given the novelty of AI technologies, people tend to know about their existence in general, but they do not always have information about how these technologies work and what specifically they do in particular applications. This is true even among software developers who do not work specifically in AI. Even though they tend to have the technical background to understand it, the state of the art in AI technologies is a pretty specialized knowledge. Those actors will have dissimilar needs of skills of knowledge, which should be measured before designing a learning curriculum. Otherwise, a literacy initiative might be disregarded as too basic, or it might be too heavy in content to be accessible for the learners.

Once those needs are identified, a data protection professional can pick **materials tailored for each audience**.

[5] See Unit 12 for a closer examination of those requirements.

- For non-technical actors, a good starting point is provided by general courses, such as the *Elements of AI* course designed to explain basic concepts without diving into their mathematical and computational implementations.
- Technical actors will likely benefit from materials that dive into AI technicalities, but they will need foundational materials in topics such as the ethics of AI and data protection obligations.
- In both cases, these general-purpose materials need to be supplemented with training materials that are specific to an organization's context. For example, part of a literacy programme could involve teaching learners how to read specific documents, such as data cards or system cards.[6]

In short, a literacy programme must spread information about how AI works, about the risks and opportunities it creates, and about the legal obligations that follow from the use of AI. This training module is designed to support the latter.

After the literacy programme is designed, it must be **kept up to date**. Given the fast pace of technical, social, and legal developments surrounding AI, many things can change quickly. Among other developments, tasks previously thought to be impossible might be solved by new AI models, the public opinion might turn against some applications of AI. Conversely, changes in the interpretation of data protection requirements might require changes to how an AI system is used. This means that not only the curriculum for AI literacy must be updated with some frequency, but that people might need to undergo regular training sessions. AI literacy is not, at least for the time being, a fire-and-forget practice.

Session 8.2. Data subject rights in the context of AI

> By the end of this session, learners will be able to **describe** some of the obstacles created by AI to the exercise of data subject rights and **discuss** whether those can be addressed by organizational measures.

One of the distinctive approaches of EU data protection law is that it grants individual rights. The data subjects whose data is processed gain certain rights they can invoke against the controllers of that processing, laid down in Articles 12–22 GDPR. Those rights are applicable to the training and deployment of AI systems whenever such practices use personal data, as discussed in the previous units. However, the peculiar technical characteristics of AI have some implications for how those rights can be exercised in practice.

[6] See Session 10.1 of this training module.

The transparency rights from Articles 13–15 GDPR will be examined more closely in Unit 11 of this training module. In the following paragraphs, we examine other data subject rights granted by the GDPR.

Restricting and objecting to processing in AI systems

The GDPR grants to data subjects two rights that allow them to affect how data controllers process their data. Under Article 18 GDPR, data subjects have the right to restrict the processing of their personal data if one of the listed conditions apply. Article 21 GDPR allows data subjects to object to processing altogether. These rights mean different things, and each has their own exceptions and conditions for application. However, their application to AI systems and models faces similar obstacles.

Those obstacles are likely to appear when the data subject attempts to exercise their right to restrict (or object to) the use of their data in the training of AI systems and models. First, a data subject might not even be aware that their data is being used for training. The transparency measures studied in Unit 11 of this training module are meant, among other things, to reduce this risk.

Additionally, a data subject might not have direct access to the organization training the model or the system. For example, a patient of **InnovaHospital** might know that the hospital is using for diagnosis an AI system based on a model developed by a third-party provider. If a patient wants to object to the use of their data for training the model, the hospital must not use that data for training (or fine-tuning the model), and it must make sure that the developer organization will not use it for training.

Things become more complicated for that data subject when the AI model is not trained on data that is specific to an organization. In that case, as discussed in Unit 6 of the training module, the organization using the model is unlikely to have control over the training process. Data subjects will need to exercise their right to restrict (or to object) against the organization training the model.

Rights to rectification and erasure

Another set of data subject rights refers to the contents of data. Data subjects can request that controllers address inaccuracies[7] or even delete[8] personal data concerning them. The conditions, exceptions, and complications related to the exercise of those rights have been extensively discussed elsewhere. But, once again, their implementation becomes more complicated when it comes to the training of an AI system.

[7] Article 16 GDPR
[8] Article 17 GDPR>

The challenge, here, is that many AI systems do not represent information in the same way as traditional computer systems. It is rarely the case that a particular piece of information is stored in a single place within the system. Instead, data about an individual is often dispersed across billions (or more) of parameters within a neural network.[9] Changing or removing that information, therefore, is not a matter as simple as making a change to an entry on a database.

Yet, data controllers remain obliged to rectify and erase personal data whenever those rights are applicable. If they fail to do so, data protection authorities can wield a variety of sanctions, including "algorithmic disgorgement," that is, the mandatory deletion of models that are not compliant with the law (Li 2022; Hutson & Winters 2024). Such a measure has not yet been deployed by data protection enforcers in the EU, and it is likely a measure of last resort against reiterated non-compliance.

Several measures have been proposed as technical and organizational alternatives to full-blown model deletion. Some of those are meant to delete data from the weights of the entire model, thus allowing its removal after the model has been trained (see, e.g., Bourtoule *et al.* 2021). Others try to make deletion feasible by changing how the model is trained. For example, the CPR technique (Golatkar *et al.* 2024) allows a model to rely not just on its core training data, but also on a private data store that can be instantly forgotten.

Those techniques are still at an early stage of development, and as such they might not be mature enough to meet all the legal requirements established in the GDPR (Cooper et al. 2024). Still, a data protection professional will need to engage with the software developers to understand whether such a technical approach is feasible in the case at hand.

The right to data portability

Article 20 GDPR equips data subjects with a right to data portability. If the outputs of an AI system qualify as personal data, a data subject has the right to request their portability. Likewise, the data subject has the right to request portability of personal data used as an input for an AI system. In both cases, the connection of the data with the AI system does not introduce any additional complications if compared with other kinds of portability.

The same cannot be said about the portability of the **weights** of an AI system based on machine learning. Given that, as discussed above, information is often spread across weights, it can be difficult to associate specific weights with a natural person. Even if

[9] See, e.g., Binns (2022).

such an identification is possible, the weights within a neural network are specific to that network's architecture. As such, they cannot be simply "plugged into" another network.

However, that transplantation of rules might be feasible in other kinds of AI systems. For example, rules codified into an expert system might be implemented in another system if the same variables are available. Therefore, a data protection professional will need to consult with the technical team to determine whether the inner workings of the model embed personal data in a format that can be ported. Future guidance by data protection authorities will bring more clarity in this regard.

Session 8.3. Automated decision-making and AI

In this session, we will continue the previous discussion on data subject rights by focusing on a specific right: the right not to be subject to an automated decision. Under Article 22(1) GDPR, a data subject has the right not to be subject to a decision that is based solely on the automated processing of personal data, if that decision has a legal or otherwise significant impact on them. Because this kind of decision is intricately connected to AI technologies, we will now spend some time on it.

This is not to say that Article 22 GDPR is applicable if and only if AI is used. Not all AI systems make significant decisions about individuals. For example, generative AI tools tend to be used to create content, while recommender systems leave the final decision to a human. Also, not all decision-making systems are powered by AI. Consider how many businesses and government organizations rely on spreadsheets to automate important processes. Even so, many large-scale applications of AI are meant to automate decisions, and this is why the risks we examined in Part I of this course often focused on decision-making tools. Hence, a training on AI cannot avoid some engagement with the provisions on automated decision-making.

Such attention is warranted because recent case law by the European Court of Justice has consolidated understanding on important aspects of this right. For one, the definition of "automated decision-making" under the GDPR is not limited to scenarios in which humans are not involved at all. This is because the court has adopted a **broad interpretation of the concept of "decision"**: it includes acts that can affect the data subject even if they do not amount to a formal decision,[10] such as the definition of a credit score that guides decisions about various aspects of an individual's life. When evaluating the applicability of Article 22 GDPR to its AI-generated outputs, an organization must therefore consider how those outputs are used within itself and by third parties.

[10] Case C-634/21, *Schufa*, ECLI:EU:C:2023:957, para. 46.

A second implication of the *Schufa* case is that the Court has confirmed that the "right not to be subject" is a **prohibition in principle**.[11] It does not require the data subject to invoke the right. Instead, it prohibits decision-making from taking place unless one of the conditions from Article 22(2) GDPR is met:

> *Paragraph 1 shall not apply if the decision:*
>
> *a. is necessary for entering into, or performance of, a contract between the data subject and a data controller;*
>
> *b. is authorised by Union or Member State law to which the controller is subject and which also lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests; or*
>
> *c. is based on the data subject's explicit consent.*

This judicial understanding is not necessarily a big shift in the application of the law, as many data protection authorities were already following this approach (Barros Vale and Zanfir-Fortuna 2022). Still, it requires caution from organizations using AI to make decisions that involve personal data.[12]

If a certain application of an AI system counts as automated decision-making for the purposes of the GDPR, its controller must ensure that one of the three legal bases above is applicable. If that is the case, the ensuing processing remains bound by the general requirements from the GDPR. Additionally, processing based on items *a* or *c* of Article 22(2) GDPR—that is, on consent or on contract—must ensure that the processing adopts measures that protect the rights, freedoms, and legitimate interests of the data subject. According to Article 22(3) GDPR, those measures must include "at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision."

This requirement can be relevant to the training and deployment of AI systems, especially those meant to replace humans in decision processes. Any organization that controls a system designed for decision-making purposes, or that repurposes an existing system for that end, should check whether the system's output counts as automated decision-making.[13] If so, they will likely need to designate specific individuals to interact with data subjects and handle their requests for human intervention, expressing their point of view, and contesting the decision.

[11] Case C-634/21, *Schufa*, para. 52.
[12] The use of special categories of personal data in decision-making is further restricted, as per Article 22(4) GDPR.
[13] On evaluating controllership, see Session 6.1 of this training module.

Contrastingly, Article 22 GDPR is **unlikely to be directly applicable to processing in the training process** of AI models and systems. This is because each processing operation during the training process does not produce a significant effect on the data subject. Nonetheless, developers of AI systems planned for decision-making should be aware that their buyers will likely need to comply with these requirements. As such, these buyers could benefit from systems that incorporate design measures that facilitate the operation of those rights.[14]

Once again, the AI Act offers extra detail to these obligations when it comes to high-risk AI systems. Under Article 14 AI Act, providers of high-risk AI systems are required to adopt technical measures that facilitate oversight. They must build the system in a way that allows the persons overseeing it:

> *(a) to properly understand the relevant capacities and limitations of the high-risk AI system and be able to duly monitor its operation, including in view of detecting and addressing anomalies, dysfunctions and unexpected performance;*
>
> *(b) to remain aware of the possible tendency of automatically relying or over-relying on the output produced by a high-risk AI system (automation bias), in particular for high-risk AI systems used to provide information or recommendations for decisions to be taken by natural persons;*
>
> *(c) to correctly interpret the high-risk AI system's output, taking into account, for example, the interpretation tools and methods available;*
>
> *(d) to decide, in any particular situation, not to use the high-risk AI system or to otherwise disregard, override or reverse the output of the high-risk AI system;*
>
> *(e) to intervene in the operation of the high-risk AI system or interrupt the system through a 'stop' button or a similar procedure that allows the system to come to a halt in a safe state.*

While such measures are not mandatory for any other AI systems, or for automated decision-making not based on AI, they provide a starting point for understanding what compliance with Article 22(3) GDPR requires.[15]

Conclusion to Unit 8

The deployment stage presents unique challenges from a data protection perspective. Though compliance with GDPR requirements demands that developers take several measures to address risks during design and development, some risks are likely to escape even their most diligent efforts. Furthermore, some risks emerge from the

[14] On some potential measures to that effect, see Unit 12 of this training module.
[15] Especially once the technical standards discussed in Session 14.1 of this training module are published.

specific context in which an AI system is put into use. This means that not all issues can be solved beforehand, and data protection professionals must be active during deployment, too.

To support them in this task, this unit has identified certain aspects that are unique to the deployment of AI systems—or at least particularly relevant in this context:

- Most people (even with a technical background) are unlikely to have a thorough understanding of how AI operates and what it does in each context. Therefore, literacy programmes can be a valuable tool for compliance.
    - Literacy training must be tailored for its audience. Some stakeholders are unlikely to need full exposure to technical detail, but they still need to grasp what AI can and cannot do. Others would benefit from looking at the system's inner workings.
    - Literacy training must be kept up to date with technical developments and changes in the social and organizational context in which the system is used.
- The technical properties of AI might create complications for the exercise of data subject rights, which organizations need to address.
    - Some of these rights will require an organization to engage with upstream providers.
    - Others require delicate trade-offs, given the limitations of current techniques.
- The broad definition of automated decision-making under the GDPR case law means that organizations might be subject to Article 22 GDPR even if there is some measure of human involvement in the loop.

These factors mean that organizations need to adopt measures and safeguards that address residual (and potentially large) risks that connect to deployment. In the next unit, we will examine some of the tools they can use for this purpose.

Prompt for reflection

The chapter highlights that exercising data subject rights, such as restriction, rectification, and erasure, can be technically complex in AI systems. Discuss potential organizational measures that could help bridge the gap between technical limitations and GDPR compliance, such as ombudsman offices or specialized teams for handling data subject requests. What role can data protection professionals play in facilitating these measures?

References

Marco Almada, 'Automated Uncertainty: A Research Agenda for Artificial Intelligence in Administrative Decisions' (2023) 16 Review of European Administrative Law 137.

Article 29 WP, 'Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679' (2018).

Article 29 WP, 'Guidelines on the Right to "Data Portability"' (2017).

Paul De Hert and others, 'The Right to Data Portability in the GDPR: Towards User-Centric Interoperability of Digital Services' (2018) 34 Computer Law & Security Review 193.

Reuben Binns, 'Analogies and Disanalogies Between Machine-Driven and Human-Driven Legal Judgement' (2022) 1 CRCL 1.

Lucas Bourtoule and others, 'Machine Unlearning' in 2021 IEEE Symposium on Security and Privacy (SP) 141.

A Feder Cooper and others, 'Machine Unlearning Doesn't Do What You Think: Lessons for Generative AI Policy, Research, and Practice', *2nd Workshop on Generative AI and Law at ICML 2024*.

Philipp Hacker, 'Sustainable AI Regulation' (2024) 61 Common Market Law Review 345.

Emmie Hine and others, 'Supporting Trustworthy AI Through Machine Unlearning' (2024) 30 Sci Eng Ethics 43.

Jevan Hutson and Ben Winters, 'America's Next "Stop Model!": Model Deletion' (2024) 8 Georgetown Law Technology Review 124.

Margot E Kaminski and Gianclaudio Malgieri, 'Algorithmic Impact Assessments under the GDPR: Producing Multi-Layered Explanations' (2021) 11 International Data Privacy Law 125.

Tiffany C Li, 'Algorithmic Destruction' (2022) 75 SMU Law Review 479.

Jakob Mökander and Maria Axente, 'Ethics-Based Auditing of Automated Decision-Making Systems: Intervention Points and Policy Implications' [2021] AI & SOCIETY.

Francesca Palmiotto, 'When Is a Decision Automated? A Taxonomy for a Fundamental Rights Analysis' (2024) 25 German Law Journal 210.

Ayush K Tarun and others, 'Fast Yet Effective Machine Unlearning' (2024) 35 IEEE Transactions on Neural Networks & Learning Systems 13046.

Sebastião Barros Vale and Gabriela Zanfir-Fortuna, 'Automated Decision-Making Under the GDPR: Practical Cases from Courts and Data Protection Authorities' (Future of Privacy Forum, May 2022).

Unit 9. Operation and Monitoring of an AI System

By the end of this unit, learners will be able to:

- **identify** data protection issues that can emerge once an AI system is put into service within an organization.
- **organize** a monitoring system to detect those issues; and
- **propose** interventions to ensure an organization's continued compliance with data protection obligations.

The operation stage of the life cycle is the goal of the entire development process: most AI systems are designed so that they can be used at some point. Given the considerable effort involving in the previous stages of the life cycle, an AI system tends to be used for a purpose, which might or not be the purpose for which it was originally designed. Either way, its use will affect the functioning of physical and virtual environments. This unit discusses what data protection obligations apply at this life cycle stage.

Those obligations are largely connected to the idea of risk. Article 25(1) GDPR obliges data controllers to consider "the risks of varying likelihood and severity for rights and freedoms of natural persons posed by the processing" when choosing which technical and organizational measures to adopt. The AI Act makes risk an even more salient factor: as discussed in Session 1.2 of this training module, the rules it applies to AI systems and models are determined based on their perceived risk. So, compliance with those requirements requires a solid understanding of what we are talking about when we are talking about risk.

Both the GDPR and the AI Act rely on an actuarial definition of risk.[1] Under such a definition, a risk is a quantity that is associated with an event that might (or not) happen. That event has a **likelihood** of happening, and if it does take place, its consequences are taken to have a measurable **severity**. The risk associated with that event is, therefore, the combination between the likelihood of the event and its severity.[2] Until a risk materializes, those values are speculative, and as such their determination suffers from the same issues of anticipation discussed in Session 9.1 of this training module. Even so, they offer an initial basis for determining how much effort should be dedicated to preventing a potentially harmful outcome.

[1] Recital 76 GDPR, Article 3(2) AI Act.
[2] Usually, calculated by multiplying one quantity by the other.

In both pieces of legislation, regulation addresses the risks to data subject rights created from the processing of data by AI systems. Yet, there are crucial differences between how the GDPR and the AI Act perceive those risks:

- The GDPR adopts a **contextual** view of risk. The AI Act follows, instead, a **top-down** approach (de Gregorio and Dunn 2022).
    - Data controllers are obliged to evaluate the risks their processing creates to data subject rights and freedoms, and then choose the technical *and organizational* measures that are best suited for eliminating or at least mitigating them.
    - In the AI Act, the legal text determines the risk categories that must be applied, leaving to the regulated actor the task of applying different sets of rules according to the risk level determined by the EU lawmaker.
- The GDPR requires data controllers to **balance** the rights at stake, including the fundamental right to data protection.[3]
    - This means, among other things, that processing must take care to avoid interfering with data subject rights if the same goals can be achieved with less interference.
    - The AI Act, instead, relies on a **satisficing** approach (Almada and Petit 2025). Any systems or models that meet the specified requirements is considered lawful, regardless of whether it just barely acceptable or if it surpasses by much the standard.
- The GDPR requires data controllers to adopt **technical and organizational measures** to address risks, whereas the AI Act largely focuses on technical measures.[4]

From these differences, one can conclude that fulfilling the AI Act's requirements is often necessary for data protection, but rarely sufficient. When using AI technologies, data controllers must not lose track of their obligations towards data subjects and their rights, even after initial deployment.

This Unit examines the risk management obligations that apply to any organization as an AI system operates. Some of those obligations fall on the actor who operates the system, but its original developer remains subject to duties both under the AI Act and data protection law. **Session 9.1** outlines how the GDPR and the AI Act oblige data controllers to manage the risks associated with the development and use of AI systems. **Session 9.2** shows how organizations are obliged to monitor issues with their issues after deployment and presents some techniques they can use to that effect. **Session**

[3] Recital 4 GDPR.
[4] See, in particular, Article 9 AI Act.

9.3 concludes the unit with a discussion of the legal obligations that bind organizations to address any post-deployment issues for the systems they are responsible for.

Session 9.1. Managing data protection risks

By the end of this session, learners will be able to **identify** risks that an organization must monitor and address once an AI system is deployed.

Under data protection law, data controllers are required to address the risks created by the AI systems, both at the moment of initial development and in any subsequent processing of personal data:

- Article 25 GDPR creates an obligation of addressing risks to data protection principles.
    - For example, if a system's accuracy degrades after its deployment, the controller must take technical and organizational measures to ensure this does not harm data subjects.
    - Such measures might include changes to the system (such as improving its model) or to its organizational context (such as removing the system form some critical applications where it would create the most risk).
- Article 32 GDPR creates an obligation of dealing with security risks.
    - For example, malicious actors might figure a way to override the safeguards adopted in a model and extract the data used for its training. If that happens, the controller must adopt measures to prevent and respond to breaches.

Those obligations apply during software development, giving origin to the obligations discussed in Unit 7 of this training module. But they also apply once the system is in service, as the risks to data protection and cybersecurity must also be faced whenever personal data is processed. Units 3 and 4 of the module gave an overview of several risks that must be considered. It is now time to discuss how an organization should weigh those risks in practice.

Both GDPR articles stipulate four factors that must be taken into account in the assessment of data protection risks: the state of the art; the cost of implementation; the nature, scope, context, and purposes of processing, and the likelihood and severity of risks. All these factors must be considered for each instance of processing, but the relative importance of each one will depend on context.

Relevant factors for risk management

Regarding the **state of the art**, the GDPR obligations mean that data controllers must consider the best practices available in the market and the current capabilities of

technology. On the one hand, this means that controllers are obliged to update their standards as technology evolves. **What is adequate today might not be tomorrow**. On the other hand, this means Articles 25 and 32 GDPR do not oblige controller to advance the state of the art. They cannot be expected to adopt innovative technical and organizational measures. However, it might be the case that a controller must refrain from processing data if no technical or organizational safeguards can reduce risk to an acceptable level.

The **cost of implementing measures** is also a relevant factor when it comes to AI. Developing AI technologies is a resource-intensive task, especially when it comes to AI models in the state of the art. Procuring AI-based tools from external sources can also be expensive, and costs are likely to increase if an organization must include extensive safeguards for data processing. As a result, the obligations of data protection and security by design would not oblige controllers to adopt measures that have an excessive cost for a very reduced mitigation of risks. But, as the enforcement of Article 25 GDPR throughout the EU shows, this does not mean that organizations can avoid adopting measures just because they are expensive. Instead, they are still obliged to adopt technical and organizational measures that reduce risks at a cost that is proportional to risk reduction.

Interpreting risk management duties

When it comes to the properties of processing and the risks it creates, evaluation will depend on the specifics of the system. For example, **DigiToys** must adopt different safeguards for the AI systems it uses in its toys and the ones it uses for data analytics, even if those systems are based in the same technologies. This is because those applications give origin to different risks. An issue with the toy itself might harm children, for example by allowing a hacker to interact directly with a child playing with a toy. Issues with data analytics, in contrast, are likely to harm the company's direct customers—for example, by exposing financial data of the parents and other people that buy those toys. Some measures that are useful for solving one type of risk, such as anonymizing financial data, might have little to offer against the other type of risk.

Given the novelty of many AI applications, it is not always easy to identify what kinds of risks must be considered for each type of processing. Still, data protection professionals can rely on a few tools to support them in that identification:

- They can **extrapolate** from existing sources of knowledge about risks of AI, such as the ones discussed in Units 3 and 4 of this training module.
- They can use **forecasting** tools such as those discussed in the next session.
- Once potential risks are evaluated, they can apply the general **guidance** offered by the EDPB in the Guidelines 4/2019, as well as materials provided by the national data protection authorities.

Despite the differences between risk framings discussed above, the AI Act can also provide some guidance for addressing the risks that AI systems can create to the rights and freedoms of data subjects. Article 9 AI Act stipulates that the providers of high-risk AI systems must adopt practices for:

- **Identifying** and analysing known and reasonably foreseeable risks that the system can pose when used in accordance with its intended purpose.
- **Estimating** and evaluating risks that may emerge both when the system is used in accordance with purpose and under conditions of reasonably foreseeable misuse.
- **Evaluating** other risks possibly arising, based on the data gathered from the post-marketing monitoring system.

Because those obligations are directed at high-risk AI systems, they are not binding in most of the data processing involving AI. Furthermore, one must be careful with the different **types** of risk that each legislation deals with, as some of the risks that are of interest for the AI Act are not risks to the fundamental rights and freedoms of a data subject.[5] What is useful here for compliance with data protection laws is, instead, the sequence of steps which an organization can follow when evaluating the risks it faces while developing or deploying an AI system.

The AI Act can also be a source of guidance regarding **which** measures to apply. Here, however, it is considerably vaguer than in the risk assessment measures. Articles 10 to 15 AI Act stipulate technical requirements that must be observed by all high-risk AI systems, but they only define the “essential elements” of those requirements. Providers of AI systems are expected to interpret these essential elements and devise their own measures for compliance.[6] Even so, the AI Act’s list of essential requirements offers a starting point that providers can adjust to their needs if they are not obliged to follow it.

Finally, the risk assessment obligations in the GDPR and the AI Act are continuing obligations. They do not end with a system’s development, or even with its initial deployment. This suggest that data controllers must consider the **timing** of their interventions to address risk. Sometimes, it might be easier to develop a workaround for a known issue in an AI system than to solve it through technical means. For example, if the **UNw** university’s AI system for forecasting student outcomes does not work well with students from non-traditional backgrounds, the university might simply create manual forecasts for those students, especially if there are few of them. However, an organization must make sure that it is actually addressing such issues at a different

[5] Which Article 25(1) GDPR obliges the data controller to protect: Guidelines 4/2019, para. 11.
[6] For some tools that can be used to guide that interpretation, see Unit 14 of this training module.

stage. Otherwise, the lack of organizational measures (or their inadequacy) might itself be a breach of the obligations on data protection and security by design.

Session 9.2. Detecting issues with AI systems

> By the end of this session, learners will be able to **explain** the various legal obligations that bind organizations to monitor data protection risks during deployment.

In the previous session, we learned that both the GDPR and the AI Act require that organizations keep track of AI-related risks throughout the entire life cycle of an AI system or model. A provider's obligations do not end when their product is commercialized but continue until it is no longer processing personal data. Likewise, the obligations of an organization deploying AI go beyond individual processing operations, encompassing all the ways it feeds personal data to an AI system and draws personal data for it. Now, it is time to examine how organizations can detect risks that can materialize after AI is deployed.

Ex ante risk detection

To some extent, detecting post-deployment risks is a matter of **anticipation**. Because the risks associated with AI are not always well-known, data controllers need to be proactive in their identification of potential risks. Otherwise, they might fail to adopt the necessary measures to address such risks and end up exposed to liability.

This means organizations must keep track of technical developments that might render their approach obsolete or **overcome existing safeguards**. For example, if somebody develops a new technique to extract personal data from medical images, some data that **InnovaHospital** previously treated as anonymous might be subject to re-identification. If that is the case, such data is now considered personal data and must be subject to appropriate safeguards.

They must also consider that **changes in the context** in which an AI system operates can affect its usefulness. Consider a scenario where if the university **UNw** starts to teach many courses in a new language, such as Chinese. If the systems it uses to predict student performance do not consider linguistic competence, they might provide an inadequate assessment of student performance. A student that has all the technical competences to succeed in a mathematics course might still struggle if they cannot understand what is being said in the classroom.

If an organization can forecast some of those changes, it might already account for them into the system and avoid the need for future change. To do so, an organization might benefit from various tools:

- **Structured techniques for prediction**, such as the Delphi method, allow organizations to combine the predictions of various experts and compensate for individual biases.
- Reports about market tendencies and **trends in technological innovation** might be a useful source of information about what is coming next in terms of technical and social developments.
- The data collected during a **system's test processes** might suggest that some aspects of the system are acceptable for now but might become a problem later. For example, one might look at a system's accuracy metrics and decide that they are acceptable for a system that makes a thousand inferences for day, but that the error levels would be unacceptable if that system were to make a million daily inferences in the future.
- Once the system is deployed, an organization can use the information it collects about its operation to extrapolate future tendencies. Coming back to the previous example, the growth in the user base might be a good indicator of whether the system usage will reach a point where a previously acceptable level of accuracy is no longer okay.

By combining those sources of information with the organization's knowledge of its context of operation, a data protection professional will be able to identify risks that they should analyse further.

Ex post risk detection

While anticipation is a valuable tool for identifying risks, a data controller cannot rely just on it, for several reasons. Sometimes **even the best forms of anticipation go wrong**: we might underestimate the likelihood of a harmful event taking place, or the extent of harm that comes out of it. For example, the development of feasible quantum computing techniques seems unlikely in the short term, but it would create all sorts of problems for current information security practices. In other cases, even robust forecast techniques might be blindsided by unexpected new developments, such as problems in datasets used to train widely used AI models. Therefore, it is likely that organizations will only learn about some of the risks of AI once they have materialized, that is, once somebody has been harmed by the use of an AI system.

It follows from this that organizations must keep track of harms that escaped their initial anticipation efforts. **This is true for at least two reasons**. Even if the harm was genuinely unforeseeable at first, it might happen again, and in that case, it is no longer unprecedented. A single wrongful diagnosis from a medical AI system might come from a bizarre set of coincidences, but understanding those circumstances would allow a hospital to prevent that error from happening repeatedly. And doing so is in its interest, as organizations remain responsible for the effects of data processing that they control.

For high-risk AI systems under the AI Act, evaluating the risks that happen is an actual obligation. Under Article 26(5) AI Act, organizations deploying high-risk AI systems must monitor the system's operation, following the instructions for use given by the system's provider. Deployers must inform any serious incidents to the provider. They are not obliged to adopt themselves any measures under the AI Act. But, as data controllers of the individual uses of the AI system,[7] those deployers will still be responsible for preventing the harms under data protection law.

Providers, in turn, are turn required to communicate with the market surveillance authorities[8] and adopt measures to fix the system. That is, a provider is required to eliminate or mitigate the possibility that the risky event will happen again. If they fail to do so, the surveillance authorities can adopt various sanctions, including fines,[9] the removal of the system from the EU market or a mandated recall.[10] If the harm stems directly from the training process of the AI system or model, they might be responsible for it under data protection law. If the harm comes from system operation, one must consider whether the provider can be classified as a data controller or processor for that processing.

How can organizations extract meaningful information from the data they collect after the system has been placed on the market? Doing so will require a mix of technical and contextual analyses:

- For the technical side of things, finding issues will likely require a **closer look at system operations**. In particular, the automated registry of system events (logging) can ensure that system behaviour is stored for subsequent analysis.
- For actually seeing the harms caused to individuals and groups, one will need to **interact with domain experts** (such as the ones operating the system) and with the **people potentially affected by the system**.

In both cases, analysis will benefit from a combination of automated tools and in-depth case studies of cases identified through automation. Once those analyses are conducted, one can start discussing how to best fix the problems found out by them.

[7] See Unit 7 of this training module.
[8] Article 73 AI Act.
[9] Article 99 AI Act.
[10] Article 74 AI Act.

Session 9.3. Addressing issues after deployment

By the end of this session, learners will be able to **explain** how organizations are obliged to address the detected issues. They will also be able to **propose** organizational strategies and practices to tackle such issues.

When it comes to risks, knowing is only half the battle. Data controllers, regardless of whether they develop AI systems or models or put those technologies to use, have various other obligations. It is not enough, for example, to detect that an attacker has found a jailbreak that allows them to change the behaviour of the AI model powering your application. You must comply with a series of legal requirements, such as notifying the data protection authority of any data breaches,[11] communicating with data subjects when the data breach results in a high risk to their rights and freedoms.[12] You must also adopting technical and organizational measures to prevent future exploitation of the jailbreak,[13] such as updating the system to close the technical exploits that enable it or even withdrawing it from service if no other measures can mitigate the risk. These obligations mean that an AI system and/or its mode of use are unlikely to remain unchanged after deployment.

In this session, we will consider some of the measures that data controllers are obliged to adopt after they have deployed their AI systems. In line with the requirements of data protection by design and security by design,[14] those measures can be technical or organizational, depending on what is best to address a specific risk in a particular context. For systems classified as high-risk under the AI Act, additional requirements apply, which must be understood considering data protection requirements. To show how that can take place, we will consider measures specific to the use of AI in automated decision-making.

Technical and organizational measures

The text of the GDPR distinguishes between two kinds of measures. **Technical measures** are technological interventions that change a system to eliminate a source of risk. Unit 12 of this training module provides examples of measures that can help with that, such as techniques for detecting biases in algorithmic decisions. The idea behind this kind of measure is that it makes the desired behaviour a part of the software's affordances. That is, the computer system will not allow a malicious user to act in a way that is contrary to data protection law, or at least make it exceedingly difficult for them to do so.

[11] Within the timeframe of Article 33 GDPR.
[12] Article 34 GDPR.
[13] Article 32 GDPR.
[14] See Unit 12 of this training module.

Organizational measures, instead, change the context in which the system operates. For example, an organization might decide to restrict access to the outputs of an AI system, to reduce the number of people that can see the personal data contained in those outputs. Those measures keep the AI system or model as it is and focus on the behaviour of the humans surrounding the technology and the context in which it operates. Both kinds of intervention can be useful for dealing with risks related to an AI system, even after that system has been deployed.

For the most part, the technical properties of an AI system are laid down during its development process. Yet, this does not mean a computer system cannot be changed afterwards. Think about the constant updates we are invited to do in the operating systems of our computers and smartphones. Those updates often add features to the systems we use or fix flaws in their security or functioning. If a developer organization detects an issue with an AI system that it already has placed on the market, it can release updates with measures that mitigate the ensuing risk.

A widespread problem with software updates is that they are not always carried out correctly. Organizations (and individual users) often postpone updates because of factors such as inconvenient timing or lack of expertise. This is why it is common to see major cybersecurity incidents that exploit vulnerabilities for which there is a known fix, such as a software update. Avoiding this kind of problem is a shared responsibility between providers and deployers of AI systems:

- Providers should make clear in the instructions for use the procedure for patching AI systems, educate deployers about the need for updates, and ideally provide support for updating.
- Deployers, on the other hand, must follow the instructions to use and keep their systems up to date.

A failure to do so is not a breach of data protection law. However, an organization that fails to update their systems to address known risks is arguably failing to adopt technical measures that can address the relevant risks. A failure to update systems may therefore lead to sanctions as a breach of the requirement of data protection (and security) by design.[15]

Organizational measures, instead, focus on the human side of the equation. Some of them relate to individual processing operations, such as establishing standard operating procedures for the use of AI systems. Others focus on preparing the individuals who will operate AI systems, as is the case for the AI literacy actions discussed in Session 8.1 of

[15] See, for examples of sanctions, Dewitte (2024).

this training model. Finally, an organization needs also to consider certain institutional channels that can support the efforts of a data protection officer:

- An organization's **customer service** can be its first line of response against risks that were not eliminated in the design process.
    - Support personnel can collect information from users about harms created by the use of an AI system, for example by processing customer complaints.
    - In some cases, it might even be feasible to grant them the power to fix those issues, for example by allowing them to undo some algorithmic decisions.
    - Even if it is not feasible to grant this kind of intervention power to customer service, communication with the affected persons is a way to ensure them that they can exercise their rights and be protected from harm.
- Internal controls, such as an **ombudsman** function, can be used to provide a critical look at current procedures and suggest how the organization can improve its use of AI.
- A robust set of **whistleblower protections** can act as a measure of last resort, allowing the people inside an organization to make sure that information about AI-related risks reach the leadership before it leads to harms in the real world.

There is no single set of organizational arrangements that will meet the requirements of the GDPR. Within a smaller organization, a clearly defined set of access procedures might harm innovation without necessarily leading to better protection of data subject rights. Requiring constant training sessions might create fatigue, making people indifferent to essential information about the risks created by AI. Guidance by a data protection professional is therefore essential for identifying the best set of arrangements for an organization developing or deploying an AI system.

Human oversight and intervention

The GDPR seldom prescribes the adoption of specific technical and organizational measures. One exception can be seen in Article 22(3) GDPR, which deals with automated decision-making. Under that provision, a data controller must adopt "suitable measures to safeguard the data subject's rights and freedoms and legitimate interests" of data subjects affected by decisions solely based on automated processing. Those measures include, at least, the right to obtain human intervention in the decision-making, to express the data subject's point of view, and to contest the decision. This requirement is frequently described as a way to keep a human "in the loop" of decision-making.

As discussed in Session 8.2 of this training module, AI is not equal to automated decision-making. On the one hand, AI systems can be used in decisions that involve

humans. For example, an AI system might suggest a few courses of action to a decision-maker, who must then choose which of those they will adept. On the other hand, automation can take place without the use of AI, as is the case in systems that use spreadsheets for risk scoring. Still, some of the applications of AI in decision-making processes can make a deployer organization responsible for following the rules in Article 22 GDPR.

For systems classified as high-risk under the AI Act, the requirements are more detailed. Any such system—even if it is just aiding rather than making the entire decision—must be subject to human oversight.[16] In particular, the provider of any such system must design it in a way that allows the persons exercising that oversight, as we discussed in Session 8.3.

The implementation of the requirements for human oversight will depend on the specifics of the system. For example, a person who oversees the functioning of a medical diagnosis system will likely need to have access to different variables and training than a person overseeing a system that is used for automated content moderation. Still, some of the requirements, such as the possibility of stopping an AI system, present a clearly defined requirement that can be implemented into a system.

The functionalities required by Article 14 AI Act are not mandatory for AI systems that are not classified as high-risk under the AI Act. Even so, the list from that article can once again be used as a starting point of measures that an organization might consider for their own systems. However, even for high-risk AI they are not sufficient. It might be the case that meaningful oversight is possible from a technical perspective but does not happen in practice. For example, a person overseeing an algorithmic system might be afraid to override its decisions if doing so will cause them to fall behind with their work or create the risk of reprisal from bosses. An organization deploying an AI system needs to take measures to ensure that the individuals exercising oversight powers can do what the law requires of them.

This obligation applies even if the organization cannot change the system itself. Even if the organization lacks the capability to make technical changes to the AI systems and models it uses, it still has control over its own internal arrangements and practices. Therefore, the legal obligation to adopt technical and organizational measures to address risks means that an organization must adapt itself for the safe use of AI.

Conclusion to Unit 9

The previous sessions have covered the issue of how one can detect and mitigate risks after an AI system has been deployed. We have discussed the legal obligations that

[16] Article 26 AI Act creates this obligation for deployers, and Article 14 AI Act obliges providers to design high-risk systems in a way that enables oversight.

create the need for ongoing risk management throughout the life cycle, what is meant by risk in the first place, and strategies for risk management during and after deployment.

A few key points emerge out of the discussion in this unit:

- The AI Act and the GDPR both tackle the risks to fundamental rights, liberties, and legitimate interests that might come out of data processing in AI. However, they do so in vastly diverse ways.
- Despite those differences, they both require **ongoing attention to risks**.
    - Tools like system logs, user feedback, and automated monitoring systems are critical to detect and evaluate risks during system operation.
    - Relevant risks might emerge at any point of the life cycle, and their characteristics can change as technologies evolve and society changes.
    - These same sources of change mean that organizations will likely need to update not just their systems but the very measures they use to detect risk.
- Risk monitoring can take place *ex ante* or *ex post*
    - *Ex ante* forecasting has its limitations, especially when it comes to the flexible uses to which AI technologies can be put. It remains a valuable tool to address some risks before they happen.
    - *Ex post* monitoring focuses on learning from harms that happened to prevent them from happening again.
- Measures for risk management
    - Both technical and organizational measures are relevant for addressing risks detected through *ex ante* and *ex post* approaches.
    - Safeguards like manual interventions or adjustments to operational workflows can act as a second line of defence.
    - The GDPR is broad when it comes to technical and organizational measures. The AI Act provides some concrete measures, which are mandatory for high-risk systems and might be useful for other technologies.

Part III of this training module will engage more deeply with some measures that emerge out of current best practices in AI development and deployment.

Prompt for reflection

InnovaHospital's deployment of AI diagnostic tools raises concerns about technical and organizational measures. Consider how different contexts (e.g., medical settings vs. educational institutions) require tailored approaches to risk management. How can organizations like **InnovaHospital** ensure that measures address the unique risks

posed by their AI systems, and how might these approaches differ from those needed by **UNw**?

References

Marco Almada and others, 'Art. 25. Data Protection by Design and by Default' in Indra Spiecker gen. Döhmann and others (eds), *General Data Protection Regulation: Article-by-article commentary* (Beck; Nomos; Hart Publishing 2023).

Frank Bannister and Regina Connolly, 'The Future Ain't What It Used to Be: Forecasting the Impact of ICT on the Public Sphere' (2020) 37 Government Information Quarterly.

Andrea Bonaccorsi and others, 'Expert Biases in Technology Foresight. Why They Are a Problem and How to Mitigate Them' (2020) 151 Technological Forecasting and Social Change 119855.

Katerina Demetzou, 'GDPR and the Concept of Risk' in Eleni Kosta and others (eds), *Privacy and Identity Management 2018* (Springer 2019).

Pierre Dewitte, 'The Many Shades of Impact Assessments: An Analysis of Data Protection by Design in the Case Law of National Supervisory Authorities' (2024) 2024 Technology and Regulation 209.

EDPB, 'Guidelines 4/2019 on Article 25 on Data Protection by Design and by Default' (European Data Protection Board, 2020).

Diana Korayim and others, 'How Big Data Analytics Can Create Competitive Advantage in High-Stake Decision Forecasting? The Mediating Role of Organizational Innovation' (2024) 199 Technological Forecasting and Social Change 123040.

Tobias Mahler, 'Between Risk Management and Proportionality: The Risk-Based Approach in the EU's Artificial Intelligence Act Proposal' in Liane Colonna and Stanley Greenstein (eds), *Nordic Yearbook of Law and Informatics 2020-2021* (2022).

Jhon Masso and others, 'Risk Management in the Software Life Cycle: A Systematic Literature Review' (2020) 71 Computer Standards & Interfaces 103431.

Jessica Newman, 'A Taxonomy of Trustworthiness for Artificial Intelligence. Connecting Properties of Trustworthiness with Risk Management and the AI Lifecycle.' (CLTC White Paper Series, January 2023).

NIST, 'AI Risk Management Framework: AI RMF (1.0)' (2023).

Jonas Schuett, 'Risk Management in the Artificial Intelligence Act' (2024) 15 European Journal of Risk Regulation 367.

Dimitrios Tsoukalas and others, ‘Technical Debt Forecasting: An Empirical Study on Open-Source Repositories’ (2020) 170 Journal of Systems and Software 110777.

Part III: Advanced Topics in AI and Data Protection

By the end of this part, learners will be able to:

- **design** assessment practices to evaluate the technical aspects of AI systems and their operation within an organization.
- **propose** technical and organizational measures to ensure that data subject rights are properly addressed throughout an AI system's life cycle.
- **devise** different information disclosure strategies in line with the legal requirements directed at each type of information recipient; and
- **evaluate** various sources of guidance that can support organizations in specifying their legal duties.

So far, this training module has offered an overview of the life cycle of AI models and systems within an organization. Each of the stages covered in Part II can give origin to various risks to the protection of personal data, which a data protection professional must address. To understand what is unique about AI in those contexts, that professional requires an understanding of the technical side of AI technologies and the vocabulary to dialogue with technical and business stakeholders, tasks that are supported by the contents of Part I of this module. Now, the remaining five units of the module will cover specific issues that are likely to span more than one stage of the life cycle.

It would not be feasible to offer an exhaustive coverage of all such issues. Each AI system or model undergoes its own life cycle, and processes that are central for a specific project might play a minor role in another. Still, current experiences with the design and implementation of AI systems and models suggest that some problems appear more often than others. This means the selection of issues for this unit is driven by two main concerns:

1. The issues covered by each unit will be relevant for most, if not all, AI systems developed or deployed within the EU.
2. Approaches to those issues offer insights that can be used to tackle other issues related to data protection in AI systems.

The knowledge and skills developed in this Part should, therefore, be applicable to a broad range of solutions, even when the specific solutions proposed here are not.

To this effect, Part III is formed by five units:

- In **Unit 10**, learners will engage with two kinds of assessments of AI systems: technical audits and impact assessments.
- **Unit 11** focuses on the information disclosure duties that organizations have with regard data subjects and public authorities, including the so-called right to an explanation.
- **Unit 12** will unpack the concept of “regulation by design”, show how it might contribute to compliance, and raise some warnings about the limits of technical interventions in promoting data protection principles.
- Drawing on those analyses, **Unit 13** zooms into a specific class of AI applications: those powered by large language models.
- Finally, **Unit 14** discusses how organizations can evaluate whether non-binding sources, such as technical standards and certification schemes, offer sufficient guidance for their compliance needs.

Mastering those topics will help learners in facing problems that cannot be confined to a specific stage of the AI life cycle.

Unit 10. Fairness and Accountability for AI

By the end of this unit, learners will be able to **exemplify** documents that can support accountability regarding AI systems and models, **sketch** the elements those documents must contain, and **plan** a documentation strategy for an organization.

Accountability is a core principle of EU data protection law. It is explicitly mentioned in Article 5(2) GDPR: a data controller is responsible for compliance with the other data protection principles, and they must be able to **demonstrate** that compliance. Furthermore, various provisions of the GDPR give effect to that principle by creating mechanisms to hold controllers to account for their processing. Article 24 GDPR establishes that controllers must adopt technical and organizational measures that allow them to demonstrate compliance with other data protection requirements. Those provisions, as is too often the case, acquire new dimensions when AI is used.

Due to AI's complexity and capacity to process vast amounts of personal data, data protection officers must ensure that these systems remain compliant with the GDPR by implementing **practices that make compliance visible and traceable**. This can be particularly relevant in AI applications in which automated decision-making processes impact individuals, such as profiling or personalized recommendations. **Regular documentation of AI-related data processing activities** is a necessary step, as it provides concrete proof of compliance efforts, allowing both internal stakeholders and external authorities to review and verify the organization's commitment to GDPR principles.

Applying accountability to AI systems presents **unique challenges**. Unlike traditional data processing, AI often involves extraordinarily complex algorithms that can operate opaquely. As a result, it can be difficult to trace precisely how personal data is processed or to understand how certain outcomes are reached. These unique characteristics create obstacles for transparency, as the underlying logic and processes in AI can be difficult for even data protection experts to interpret.

This opacity is problematic from a GDPR perspective, as accountability requires a level of transparency and demonstrable control over data flows and decision-making processes. For data protection officers, this means conducting thorough assessments of how AI systems handle data and analysing the decision-making processes these systems employ. Such assessments enable organizations to meet GDPR's accountability requirements by ensuring that they understand and can explain how AI systems operate, thereby facilitating both compliance and transparency.

One necessary step to carrying out this kind of assessment is the reduction of the various forms of opacity surrounding an AI system. In Session 11.3 of this training module, we examine technical approaches towards the **explanation** of AI systems, that is, techniques that allow one to understand the technical factors that guide a system's decision processes. But, as we discussed in Session 4 of the module, opacity is not solely a technical problem: there are also legal factors that prevent the release of information about an AI system. And sometimes technical complexity is even instrumentalized to prevent the release of information about an organization's practices.[1] As such, those technical measures for transparency need to be supported by accountability measures that ensure an organization's decisions on how and what to disclose can be evaluated.

We now examine three issues that are relevant for accountability when AI systems are used to process personal data. **Session 10.1** discusses how various kinds of software documentation can assist organizations in demonstrating compliance with data protection requirements. **Session 10.2** deals with a specific kind of document that is sometimes required by the GDPR: the data protection impact assessment. Finally, **Session 10.3** discusses how data controllers can responsibly pursue fairness in AI systems.

Session 10.1. Documenting technical decisions

> By the end of this session, learners will be able to **distinguish** between the various roles of technical documentation and **map** elements that are need for documentation to support accountability.

Large software projects are often accompanied by various kinds of technical documents. Those documents are drawn up in response to several needs, such as:

- **Registering** and explaining strategic decisions for later implementation.
- **Supplying** technical detail about what has been done within a system, to facilitate future updates and maintenance actions.
- **Guiding** the potential future users on how they operate an AI system; or
- **Demonstrating** how the system complies with software requirements.

Those needs often require vastly distinct types of documents. The level of detail that is adequate for a software developer learning about a system is likely to be too complex for an operator who just needs to understand what the system does and how to use it. Yet, each of those documents can be relevant for different data protection tasks.

[1] See, among others, Busuioc et al. (2023).

In line with its technology-neutral approach,[2] the GDPR mostly refrains from prescribing specific types of documents. In some cases, as we will soon see in Session 10.2, the proper deployment of an AI system might require a data protection impact assessment. However, the GDPR focuses on expressing the contents that must be supplied and not the form of expression. Article 15 GDPR, for instance, allows data subjects to request some information from data controllers, while Article 24(1) GDPR obliges controllers to be able to show that processing is in conformity with the GDPR.

It is true that documentation creates some frictions with development processes. They demand additional effort to draw up and maintain updated documentation, and the sheer volume of documents relating to a large AI system can be intimidating. Because those efforts are often seen as having limited returns, one of the key tenets of agile software development is the idea that working software is more important than comprehensive documentation. This commandment does not mean that documentation should not exist. But it suggests the need to minimize written records to what is essential for business reasons, including compliance with the law.

The compliance roles of software documents

One of the challenges organizations face in determining what documents are *essential* is that there is no closed list of such documents. This is because the value of documented information varies with context. A piece of information that is useless for understanding the impact of a system used in social media might make all the difference for assessing whether a medical diagnosing system works as intended. Some types of documents are mandated by law, such as those demanded by sector-specific law. Others emerge as industry standards, as technical experts deem some kinds of information to be essential for their work and for accountability. This session cannot offer an exhaustive list of such documents, but it will introduce some that are deemed to be useful for AI governance.

The first type of documents that can come in handy for an organization relates to the **decisions** it makes during the software life cycle. Any organization that develops an AI system makes various choices throughout the development process: what algorithms should be used? What data is relevant for training this AI system? How should we test the completed AI system? Likewise, a deployer of an AI system must make choices such as determining which AI system to use and how to use it. In both cases, the choices will shape how personal data is processed. As a result, Article 24 GDPR entails that the organizations must be able to demonstrate those choices are made in compliance with data protection law.

[2] Recital 15 GDPR.

By documenting the process behind those choices, the actual choices made, and how they are implemented, an organization can demonstrate its due diligence regarding the numerous factors highlighted in Part II of this training module. Organizations providing systems classified as high-risk under the AI Act are obliged to provide this kind of information,[3] covering at least the criteria flagged in Annex IV AI Act. Any other controllers are not bound by this requirement. Still, they should consider documenting those decisions, especially those that create (or address) more risk.

The second type of documentation that is relevant for AI systems refers to a system's **instructions for use**. When an organization provider an AI system, it makes certain assumptions about the purposes for which their system might be used and how somebody might use the system for those ends. Even if a provider does its part in anticipating risks,[4] the system might still cause harm if the deployers ignore the measures and safeguards put in place to address risk. For example, if the university **UNw** inputs personal data about students in a public chatbot that uses that data for training, some of that personal data might become accessible to other users of the chatbot. Following the instructions for use is an organizational measure to mitigate risk.

Finally, an organization might want to document the **results** of system operation:

- For a provider, this might mean keeping a paper trail of the software testing it conducts[5] and the results of any audits,[6] as well as any bug reports received from its customers afterwards.
- For a deployer, responsibility entails keeping track of what happens during system operation, to contact providers, affected parties, and the relevant authorities in case of harm.

As previously discussed, the AI Act creates specific requirements in this regard for high-risk AI systems.[7] But the responsibility to follow results is already present in data protection law. So, it applies regardless of risk level.

Best practices in AI system documentation

Documentation does not exist for the sole purpose of compliance. It also plays a variety of other roles in software. Some types of documents help software developers in upgrading and maintaining existent systems, while others help prospective buyers make sense of the tool. In some applications, there might even be an interest in making information about the system accessible to the public. For example, the use of AI in public-facing applications might be made more legitimate by making clear to the public

[3] See the documentation requirement in Article 11 AI Act.
[4] See Unit 9 of this training module.
[5] See Session 7.2 of this training module.
[6] See Session 7.3 of this training module.
[7] See Session 9.2 of this training module.

the role of the AI system. Accordingly, we will now consider some **best practices** for documentation.

One best practice is to ensure that documentation is **comprehensive and structured**. This means clearly defining sections within documents to address several aspects of the AI system, such as data sources, processing methods, model performance, and ethical considerations. By adopting a standardized structure, organizations can facilitate easier navigation and understanding for various audiences. For instance, technical teams may require in-depth details about algorithms and data processing techniques, while executive management may need a high-level overview that focuses on compliance, risk management, and strategic implications.

On a related note, it is crucial to **tailor the language and content** of the documentation to the specific audience. For technical audiences, documentation should include precise terminology and detailed descriptions of algorithms, data processing methods, and system architecture. In contrast, documentation aimed at non-technical stakeholders, such as compliance officers or executives, should focus on implications for data protection, compliance status, and risk assessments, avoiding overly technical jargon. This approach ensures that all stakeholders can access the information relevant to their roles and responsibilities, enhancing overall understanding and engagement with the AI system.

Another important aspect is to maintain **up-to-date documentation**. AI systems can evolve rapidly, with models being updated or new data sources introduced. Organizations should implement processes for regularly reviewing and revising documentation to reflect these changes accurately. This practice not only aids in compliance with the GDPR’s accountability requirements but also supports internal audits and assessments, as outdated information can lead to misunderstandings and increased compliance risks.

Finally, organizations should include a section on **ethical considerations and potential biases** in their documentation. This part should address how the AI system is designed to mitigate bias, the diversity of the training data, and any measures taken to ensure fairness and transparency in automated decisions. By documenting these aspects, organizations demonstrate their commitment to ethical AI practices and provide data protection officers with the necessary insights to address potential risks related to data subjects' rights and freedoms.

Session 10.2. Varieties of impact assessment for AI

By the end of this session, learners will be able to **distinguish** between various kinds of impact assessment report that are associated with AI systems and **identify** when each type of report is legally required.

A data protection impact assessment (DPIA), as it names suggests, is an evaluation carried out by a data controller before they carry out certain forms of high-risk data processing.[8] As defined in the GDPR, a DPIA is required whenever the nature, scope, context, and purposes of processing suggest it is likely to result in a high risk to the rights and freedoms of natural persons. That same provision highlights that the use of "new technologies" is likely to trigger the need for an impact assessment. This session, accordingly, discusses when a DPIA is required for AI systems and what should be contained in that assessment.

In particular, a DPIA is required when there is:[9]

1. A systematic and extensive evaluation of personal aspects relating to natural personals, which offers the base for decisions that produce legal effects (or similarly significant effects) for the concerned legal person.
2. Large-scale processing of special categories of personal data[10] or personal data relating to criminal convictions and offences.[11]
3. A systematic monitoring of a publicly accessible area on a large scale.[12]

Some of those hypotheses are also present in the AI Act's list of high-risk AI systems under Annex III. This does not mean, however, that a DPIA is only needed for systems classified as high-risk under the AI Act. After all, the GDPR uses the risks of **processing** as the relevant criterion for determining the need for a DPIA, whereas the AI Act is concerned with the **technical system** as a whole. Often, systems that are not particularly risky from a technical standpoint might nonetheless create problems when (mis)used in sensitive contexts, as shown, for example, by various spreadsheets used for assessing the risk of benefits fraud in Dutch municipalities. Technical complexity is a complicating factor when choosing the measures that need to be applied, but the lack of complexity is not necessarily a sign that an application does not create data protection risks.

[8] Article 35(1) GDPR.
[9] Article 35(3) GDPR.
[10] Article 9(1) GDPR.
[11] Article 10 GDPR.
[12] Learners working in law enforcement should also consider Article 5(1)(h) AI Act.

DPIA before the deployment of an AI system

For deployers, the first step is to self-assess whether the AI system's processing of personal data constitutes a high risk to the rights and freedoms of natural persons. This is likely to be the case for a system classified as high-risk under the proposed AI Act, as the Act's risk classification is based on the impact of AI systems in fundamental rights. For example, consider a scenario where **InnovaHospital** decides to use AI in a medical device that fall into the most strictly regulated classes of the Medical Devices Regulation. The use of an inadequate AI system can create risks to (among others) the right to health of the patients exposed to the device. As such, the system not only meets the AI Act's definition of high risk under Article 6(1). It is also likely to create the kind of high risk that demands a DPIA under the GDPR.

However, the DPIA's risk requirement might be met even if a system is not classified as high-risk under the AI Act. For instance, AI applications in tax administration could pose significant privacy risks due to the potential for affecting individuals' legal rights. As such, they are likely to require a data protection assessment, even if the use of AI in tax is explicitly excluded from the AI Act's definition of "law enforcement." This example illustrates that the AI Act can offer a guideline to the application of Article 35 GDPR, but it does not replace a data controller's careful evaluation of the context.

When conducting DPIAs for high-risk AI systems, deployers must integrate information provided by the developer.[13] At the very least, this means deployers should use the developer-provided instructions for use, which often outline the AI system's operational parameters, limitations, and specific conditions for safe use. This information is critical for understanding how the system might impact data subjects and for identifying appropriate safeguards that align with GDPR's accountability standards.

DPIA during the AI development process

The AI Act does not create a similar obligation for providers. That is, organizations developing high-risk AI systems are not obliged to incorporate into the DPIA any information they obtain from upstream providers. Nonetheless, those organizations are likely to need to carry out a DPIA themselves. This is because any processing of personal data in the training process is likely to meet the requirements from Article 35(1) GDPR:

- If such processing occurs, its goal is to create a system that, by definition, poses a high risk to the rights and freedoms of the natural persons affected by the system.[14] As such, the risk criterion is likely to be met for the training process.

[13] As mandated by Article 26 AI Act.
[14] Which are not necessarily the same persons whose data is being processed.

- The training of an AI system is, at least for the time being, an operation involving novel technologies. In the future, when techniques for training AI mature enough, this might no longer be the case.
- Many AI systems are trained with the use of personal data. Whenever that is the case, the training might fall within the scope of the GDPR.[15]

Considering these factors, a provider developing an AI system will likely need to conduct a DPIA before they can commercialize that system or put into service. As they do so, they might benefit from the information made available in their own technical documentation. Additionally, they might want to use information obtained from their own providers. For example, **InnovaHospital** might want to refer to ChatGPT's documentation as it assesses a chatbot that uses this model. Doing so will allow an organization to see the bigger picture of risks associated with a system.

Likewise, developers of general-purpose AI models trained on personal data would do well to carry out a DPIA before placing their products on the market. If a general-purpose model has systemic risk,[16] it has the potential to impact fundamental rights at a large scale. Therefore, its training is a textbook example of the kind of risky processing with novel technologies covered by Article 35(1) GDPR. Even for models that fall short of the technical threshold for systemic risk, the level of risk might still be high enough. This is the case especially if a model relies on special categories of personal data.[17] Hence, a DPIA is not an obligation just for the organizations deploying AI systems and models, but also for the ones creating them.

Other impact assessment reports

In the broader context of corporate social responsibility, businesses are often encouraged (by industry associations, consumers, and other stakeholders) to carry out human rights impact assessments (HRIA) of their AI solutions. In a more binding fashion, Article 27 AI Act obliges some deployers of high-risk AI systems to carry out a fundamental rights impact assessment (FRIA). Because these assessments require an extensive evaluation of the AI system in question, completing them demands resources. In the rest of this session, we will examine those requirements.

A FRIA is required under the AI Act **before the initial deployment** of certain high-risk AI systems. As specified in Article 27(1) AI Act, a FRIA is required if the deployer of the high-risk AI system is governed by public law, or if it is a private entity carrying out public services. For example, the university **UNw** would likely be required to carry out a

[15] See Unit 6 of this training module.
[16] According to the criteria in Article 51 AI Act.
[17] See Article 35(3) GDPR.

FRIA for its high-risk AI, as a public university. This kind of impact assessment is also required of two types of private actors carrying out private functions:

1. Those using AI for evaluating the creditworthiness of natural persons or establish their credit score (except systems used for detecting financial fraud).
2. Those using AI for risk assessment and pricing in life and health insurance.

Because those two applications are themselves listed as high-risk in Annex III AI Act, any AI system used for those purposes requires a FRIA.

If a FRIA is needed, it must include certain kinds of information:

(a) a description of the deployer's processes in which the high-risk AI system will be used in line with its intended purpose;

(b) a description of the period of time within which, and the frequency with which, each high-risk AI system is intended to be used;

(c) the categories of natural persons and groups likely to be affected by its use in the specific context;

(d) the specific risks of harm likely to have an impact on the categories of natural persons or groups of persons identified pursuant to point (c) of this paragraph, taking into account the information given by the provider pursuant to Article 13;

(e) a description of the implementation of human oversight measures, according to the instructions for use;

(f) the measures to be taken in the case of the materialisation of those risks, including the arrangements for internal governance and complaint mechanisms.

A careful read of the list above suggests a considerable overlap with the impacts to fundamental rights covered by a DPIA. To a lesser extent, the same can be said of HRIAs. While there is no single list of elements required by a HRIA, as the methodologies are chosen based on business requirements, they all cover the impact of AI systems on human rights, which include the fundamental rights outlined above.

It might be possible in some cases to offer a single report that covers all the points required by data protection law and those human rights-focused instruments. Even if that is not the case, much of the work done in the elaboration of the DPIA will be relevant for drafting those reports. Hence, the DPIA, the FRIA, and the myriad forms of HRIA should not be seen as competitors, but as allies in the shared goal of producing trustworthy AI.

Session 10.3. Pursuing fairness in AI technologies

By the end of this session, learners will be able to **outline** a data protection impact assessment for an AI system or model and **combine** that assessment with other assessments that might be required by EU law.

Fairness is a critical concept for AI. Many of the problematic uses of AI technologies we discussed in Unit 4 can be ultimately traced to the unfair impact that the use of AI has to individuals in those circumstances. Furthermore, Article 5 GDPR establishes fairness as one of the guiding principles of personal data processing.[18] For all the widespread agreement that fairness matters, it can be exceedingly difficult to pin down how exactly it matters and what we should do about it. In this session, we will examine how to find the substance of the legal duties of fairness under the GDPR.

We will not examine here the definitions of fairness metrics. Learners interested in those technical details would be well-advised to consult the companion training module.[19] This session discusses, instead, some factors that data protection professionals must consider when helping technical actors in the selection of metrics that are relevant for particular cases.

Different conceptions of fairness

When it comes to fairness in AI systems, we must deal with the overlaps and conflicts between different conceptions of fairness. From the perspective of data protection law, Article 8(2) of the EU Charter of Fundamental Rights establishes that everyone's personal data must be processed fairly. This principle, as discussed in Session 6.3 of this training module, can ultimately be interpreted as a requirement of trust (Roßnagel and Richter 2023, p. 268): if one is processing an individual's personal data, they must do so in a way that warrants the trust of the data subject.

That is, it is not enough that an individual **trusts** the data controller, as they might do so for the wrong reasons. The conditions of processing must be such that the data subject's rights and interests are not disturbed excessively or without justification. What that means in practice is not determined by data protection law itself, but by broader considerations, such as those relating to EU discrimination law (Weerts et al. 2023).

This view of fairness bears some relationship to how fairness is perceived in computer science but is ultimately distinct from it. From a computer scientist's perspective, the legal—and ultimately philosophical—challenges of fairness become the technical problem of algorithmic fairness. A vast body of research has been dedicated to this

[18] On that, see Session 6.3 of this training module.
[19] Enrico Glerean, *Elements of Secure AI Systems.*

problem over the past few years, which focuses evaluating on whether and how a decision made by an AI system can treat different data subjects equally.

Technical research on algorithmic fairness requires two separate tasks. One needs to propose a metric that formalizes what it means to be unfair, defining the concept in a way that can be given a mathematical treatment. Based on that formulation, it becomes possible to measure how unfairness (under that definition) takes place in a concrete context and evaluate whether proposed technical interventions increase or reduce that unfairness (Weinberg 2022). By implementing such techniques, providers and deployers of AI systems can increase the fairness of their data processing operations, in line with the spirit of the law.

However, the difference between legal and technical conceptions of the fairness problem has practical implications. As recent interdisciplinary studies point out (such as Wachter et al. 2021, Weerts et al. 2023), EU law understands fairness and non-discrimination in a way that is both highly contextual and somewhat different from how those concepts are treated in the United States, where a considerable part of technical research on AI fairness takes place. The contextual character of fairness makes it difficult to express in formal terms that can be expressed in a computer, precluding full automation of fairness checks. The legal differences between the EU and the US, in turn, mean that many of the metrics proposed for algorithmic fairness do not tackle the same problems required by EU law. As a result, one should be careful when using fairness metrics as a tool to evaluate a system.

This is not to say that algorithmic fairness studies are of no value from the perspective of data protection compliance. To the contrary: some of these metrics capture important aspects of the phenomenon, and so they suggest ways to make a system fairer. If one is aware of the limitations of the tools, it should be possible to use them in a fruitful way. Additionally, the fairness-promoting measures suggested by that body of research might be adapted to better suit EU law requirements. Indeed, the studies mentioned above are part of a growing literature that suggests how to use metrics that are thought for the European context.

Technical limits of algorithmic fairness

Beyond legal problems, the pursuit of algorithmic fairness can also be criticized in technical grounds. The first, and sometimes most salient, critique is that algorithmic fairness might be a problem that is impossible to solve. In an early paper in the field, Jon Kleinberg and co-authors (2017) showed that, except in some very narrow cases, it was impossible to find a solution that could satisfy at the same time some of the most accepted definitions of algorithmic fairness. Given that such metrics are thought to describe some aspect of what fairness really means, this result suggests that

algorithmic fairness cannot be achieved and the best we can hope for is some kind of trade-off.

The existence impossibility result has not stopped research in algorithmic fairness. In fact, researchers have devised numerous ways to try and square the circle of algorithmic fairness. Some (e.g. Beigang 2023) have proposed modifications to the criteria, changing their formulation so that they are no longer incompatible but still capture the underlying intuitions about what fairness means. Others have proposed that one or more of the alternatives considered by Kleinberg and co-authors must be abandoned, potentially in favour of metrics that capture what fairness is *truly* about. Still, what these metrics pursue is the fair treatment of individuals *vis-à-vis* others, not the fair processing of an individual's personal data, which is what data protection law is concerned about.

Future legal guidance might help data controllers in choosing how to approach this problem in practice. Until such guidance comes along—for example, in the form of harmonized technical standards that deal with fairness[20]—providers and deployers of AI technologies should be aware that the choice of fairness metrics can be controversial.

Another problem that is usually raised about algorithmic fairness is its unitary approach. That is, algorithmic fairness research often tries to find a single set of conditions that will tell us whether something is fair (Beigang 2022). This is not necessarily a good reflection of the world, as people might have well-grounded but still diverging criteria of what fairness requires. In fact, one might say that many political disputes are precisely disputes about what is fair. So, the decision to follow one specific view of fairness might always be questioned by those for whom that view is unacceptable.

In addition, it has been suggested that viewing fairness as a single set of criteria blurs important distinctions. In a 2022 article, Fabian Beigang argues that unfairness can emerge in two different moments when AI is used in decision-making processes. First, the prediction generated by the AI system itself might be unfair, as is the case if the system produces discriminatory outputs. Second, unfair treatment might happen when the algorithmic output is used to allocate resources. For example, an unbiased facial recognition model might still be used for supporting discriminatory decision-making, such as policies that segregate people from a specific ethnic background. By looking at those two issues separately, an organization might have more clarity about the fairness issues that its use of AI can create.

Conclusion to Unit 10

We can summarize our previous discussion as follows:

[20] See Session 14.1 of this training module.

- Technical documentation provides **organizational memory**, as it registers what decisions were made in the development process, the alternatives that were considered, and the outcome of debates.
    - There are **several types of technical documents**, each aimed at a certain audience that requires a particular level of detail.
    - Some **best practices** in software documentation can be used to ensure that the documents are sufficiently informative.
    - When elaborating those documents, one should take care to store decisions **about** data protection requirements, as well as those that might be relevant to understand issues later.
- Many uses of AI technologies with personal data might require a data protection impact assessment, but not all of them.
    - DPIAs might be needed both during the development and for the deployment of an AI system. At each point, they might require distinct types of information, but the overall aim is the same: paying proper attention to how the use of AI can impact the rights, liberties, and interests of others.
    - **DPIAs coexist with several types of impact assessments**, such as those related to human and fundamental rights. When there is a substantive overlap between reporting requirements, organizations might avoid rework by integrating the contents of different reports.
- **Fairness is a complex concept**, which is not exhausted by the technical formulation of algorithmic fairness.
    - Some technical definitions of fairness are incompatible with one another.
    - Some legal aspects of fairness are not well represented in formal representations.
    - Nonetheless, algorithmic fairness approaches can be useful for legal compliance if one pays attention to their limits.

In this unit, we have examined some of the mechanisms data protection law and the AI Act utilize to stimulate the fair and accountable use of AI technologies. Documents, such as the technical software documentation and the various kinds of impact assessments discussed above, can provide a paper trail that is fundamental for justifying and evaluating why a system functions in a certain way. Part of that assessment is likely to deal with whether the design and use of the system reflect the GDPR principle of fair processing, and a gap in accountability might itself be something that makes processing unfair. Therefore, fairness and accountability are intricately connected in AI systems.

Prompt for reflection

Data Protection Impact Assessments (DPIAs) are a core tool for assessing risks under GDPR, but fairness is often a less tangible concept that is hard to measure. Reflect on how DPIAs can incorporate fairness considerations effectively.

References

Marco Almada and Nicolas Petit, 'The EU AI Act: Between the Rock of Product Safety and the Hard Place of Fundamental Rights' (2025) 62 Common Market Law Review.

Jens Ambrock and Moritz Karg, 'Art. 35. Data Protection Impact Assessment' in Indra Spiecker gen. Döhmann and others (eds), *General Data Protection Regulation: Article-by-article commentary* (Beck; Nomos; Hart Publishing 2023).

Christoph Bartneck and others, 'Trust and Fairness in AI Systems' in Christoph Bartneck and others (eds), *An Introduction to Ethics in Robotics and AI* (Springer 2021).

Fabian Beigang, 'On the Advantages of Distinguishing Between Predictive and Allocative Fairness in Algorithmic Decision-Making' (2022) 32 Minds and Machines 655.

Fabian Beigang, 'Reconciling Algorithmic Fairness Criteria' (2023) 51 Philosophy & Public Affairs 166.

Madalina Busuioc, Deirdre Curtin and Marco Almada, 'Reclaiming Transparency: Contesting the Logics of Secrecy within the AI Act' (2023) 2 European Law Open 79.

Giovanni de Gregorio and Pietro Dunn, 'The European risk-based approaches: Connecting constitutional dots in the digital age' (2022) 59 Common Market Law Review 473.

Margot E Kaminski, 'Regulating the Risks of AI' (2023) 103 Boston University Law Review 1347.

Jon Kleinberg, Sendhil Mullainathan and Manish Raghavan, 'Inherent Trade-Offs in the Fair Determination of Risk Scores', *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)* (Schloss Dagstuhl--Leibniz-Zentrum für Informatik 2017).

Eleni Kosta, 'Article 35. Data Protection Impact Assessment' in Christopher Kuner and others (eds), *The EU General Data Protection Regulation (GDPR): A Commentary* (Oxford University Press 2020).

Alessandro Mantelero, *Beyond Data: Human Rights, Ethical and Social Impact Assessment in AI* (Springer Nature 2022).

Claudio Novelli and others, 'AI Risk Assessment: A Scenario-Based, Proportional Methodology for the AI Act' (2024) 3 Digital Society 13.

Alexander Roßnagel and Philipp Richter, ‘Art. 5. Principles relating to processing of personal data’ in Indra Spiecker gen. Döhmann and others (eds), *General Data Protection Regulation: Article-by-article commentary* (Beck; Nomos; Hart Publishing 2023).

Jonas Schuett, ‘Risk Management in the Artificial Intelligence Act’ [2023] European Journal of Risk Regulation FirstView.

Sandra Wachter, Brent Mittelstadt and Chris Russell, ‘Why Fairness Cannot Be Automated: Bridging the Gap between EU Non-Discrimination Law and AI’ (2021) 41 Computer Law & Security Review 105567.

Alina Wernick, ‘Impact Assessment as a Legal Design Pattern—A “Timeless Way” of Managing Future Risks?’ (2024) 3 Digital Society 29.

Hilde Weerts and others, ‘Algorithmic Unfairness through the Lens of EU Non-Discrimination Law: Or Why the Law Is Not a Decision Tree’, *2023 ACM Conference on Fairness, Accountability, and Transparency* (ACM 2023).

Lindsay Weinberg, ‘Rethinking Fairness: An Interdisciplinary Survey of Critiques of Hegemonic ML Fairness Approaches’ (2022) 74 Journal of Artificial Intelligence Research 75.

Unit 11. Transparency towards Stakeholders

By the end of this unit, learners will be able to:

- **distinguish** between the different stakeholders to which organizations are obliged to provide information about their use of AI.
- **break down** the different informational needs of those stakeholders and the types of information that must be provided; and
- **propose** compliance approaches that ensure the information being provided is fit for purpose.

Both the GDPR and the AI Act require the providers and deployers of AI systems to disclose various kinds of information to stakeholders. The specific kinds of information that must be disclosed under each legal instrument will depend on the legal classification given to each actor. For example, a business that deploys an AI system will likely be classified as a data controller under the GDPR for the data processed by that system, and as such it will be subject to certain transparency requirements. At the same time, it will likely be classified as a deployer, and thus subject to the requirements that apply to this kind of actor.[1] Still, an organization cannot disclosure information if it has not access to it in the first place.

Regulatory authorities have considerable powers to request information from the regulated actors.[2] The persons affected by the use of an AI system also have some rights to receive information about its operation.[3] And the general public has a limited right to obtain information about some kinds of AI systems, as Article 49 AI Act mandates the registration of high-risk AI systems into a publicly available database. Additionally, Article 50 AI Act mandates that providers and deployers of AI systems disclose when they are using AI for some applications, such as interaction with humans or the generation of artificial content, especially when it cannot be distinguished from authentic. As these examples show, the regulation of AI in the EU gives a high value to diverse kinds of transparency.

This unit discusses three forms of disclosure that are both necessary and complicated in contexts involving AI:

- **Session 11.1** discusses legal duties to disclose information to regulatory authorities.

[1] For high-risk AI systems, Articles 26 and 27 AI Act are particularly relevant here.
[2] Articles 58 GDPR, 74 AI Act.
[3] See Session 8.2 of this training module.

- **Session 11.2** examines whether and how the developer of an AI system or model must disclose information about it to downstream developers who might want to use it in their own systems.
- **Session 11.3** evaluates current approaches for technical AI transparency from the perspective of whether they can support compliance with data protection duties.

Session 11.1. Disclosure duties towards public bodies

> By the end of this session, learners will be able to **adapt** technical and organizational practices regarding information to ensure that an organization can provide meaningful information upon request by data protection authorities.

A key element of the GDPR's enforcement systems is that regulators have substantial investigative and corrective powers. Under Article 58 GDPR, a supervisory authority can order controllers and processors to provide "any information it requires for the performance of its tasks", as well as to carry out audits. Those and other investigative powers remain in force when AI systems are used to process personal data. They also apply when personal data is used in the training of AI systems and models. As such, data controllers and processors need to store and keep up to date the kind of information that a DPA will need if it needs to investigate the AI system or model in question.

When it comes to high-risk AI systems and general-purpose AI models, the AI Act adds more details both to the kind of information that needs to be stored and to the powers of supervisory authorities. Article 74 AI Act grants to market surveillance authorities have the power to obtain access to documentation, data sets, and even the source code of high-risk AI systems in certain cases, and as we shall see below, providers and deployers of high-risk systems and general-purpose AI models are required to keep some information for the purpose of compliance. Therefore, the AI Act reinforces the GDPR's overall approach of obliging organizations to provide extensive support to regulators in their supervisory duties.

Confidentiality as a condition and a limit for disclosure

Because such information is extensive, its disclosure is potentially disadvantageous for organizations. One concern is that, if that information were to become public, people would be able to subvert or otherwise manipulate the AI systems or models. For example, if **UNw** adopts an algorithm to detect cheating in university examinations, a student that knows how that algorithm works might devise a means to avoid detection. This risk of gaming is often invoked by public sector authorities as a reason some

aspects of algorithm design in domains such as fraud risk assessment cannot be made public.

The disclosure of information relating to an AI system or model can have consequences even if it is not used against the system or model itself. For example, a pricing model developed for a business will likely be developed and trained from information that is available to that business and consider elements of its commercial strategy. A competitor that replicates the model might benefit from those insights and the business's technical work at a fraction of the cost. It might also be able to extract business secrets from the model. To avoid such risks, public and private organization both use the various strategies for opacity discussed in Session 4.3 of this training module.

Acknowledging those concerns, both the GDPR and the AI Act feature mechanisms to balance the regulators' need for information and the data controllers' need for secrecy. The GDPR stipulates that the exercise of regulatory powers by data protection authorities must be accompanied by "appropriate safeguards", which include effective judicial remedy.[4] More specifically, it binds the staff members of those authority to a duty of **professional secrecy** with regard to confidential information they receive during their work, which continues to apply even after the end of the staff member's term of office.[5] The AI Act likewise requires that all regulatory authorities observe a duty of confidentiality, with special attention to the protection of intellectual property rights, trade secrets, and public and national security interests.[6] The result is a system in which the information shared by public and private controllers and processors is protected against leaks from the DPA.

The other side to this elevated level of protection is that data controllers and processors are expected to be forthright when they release information to the supervisory authority. A failure to keep the information that is necessary to understand how a system processes data, or to supply it to authorities on request, can itself lead to sanctions, in addition to any sanctions that might come out of a potential GDPR breach. In the rest of this session, we will discuss what kinds of information must be provided in this context.

Information that must be made available to the authorities

In Part II of this training module, we covered a variety of data protection issues that can emerge from the development and use of AI technologies. Addressing those risks falls within the remit of data protection supervisory authorities. This means that the authorities will need access to information that allows them to identify how a particular data processing operation can harm data subject rights. It will also need the contextual

[4] Article 58(4) GDPR.
[5] Article 54(2) GDPR.
[6] Article 78 AI Act.

detail to understand what kind of technical intervention is desirable: should the DPA order the data controller to pursue a technical fix? Mandate certain organizational measures? Or stipulate that the system cannot be salvaged at all? To arrive at those important decisions, a supervisory authority needs to consider the issues that can emerge at each step of an AI technology's life cycle.

The first thing that must be said about those requirements is that they do not mandate any specific type of document. If an organization provides the information needed by the supervisory authority, it can do so in any form. Meeting the GDPR's requirements, or even the AI Act's, does not mean that an organization needs to forsake agile software practices for a waterfall model. What it *does* require is that organizations take care regarding the substance and the validity of the information contained in the documents.

Regarding validity, an organization must make sure that the documents reflect the version of the system that it uses. Otherwise, a comprehensive documentation might be even misleading.

One type of documentation issue an organization wants to avoid is a failure to describe safeguards that are in place. Consider a scenario in which **DigiToys** fails to mention that they adopted a tool for anonymizing some of the data they collect from children. This omission creates issues for the company, which will be expected to adopt safeguards for data that is not actually personal data. It also prevents adequate scrutiny of the system, as it does not provide information that is needed to evaluate whether the anonymization techniques are suitable for their purpose. The result is a scenario in which the documents offer an incomplete, and perhaps misleading, guide to the system.

Documentation might also be misleading if it is not accurate regarding the details of the system. For example, suppose the documents for one of **InnovaHospital**'s automated diagnosis tools fail to mention a change to the model used to power the system's functionalities. If that happens, the data protection authority might end up requiring that the organization adopt safeguards that are not relevant for the current model. Keeping documentation up to date is not just an exercise of checkbox compliance, but something that can help organizations in understanding the technical and legal risks they face.

As for the contents of the documents, they will depend on the techniques being used to produce an AI system or model, and in the context in which that object is sold or used. An organization would do well to write down the analyses it conducts in the context of the various stages covered by its training module: what issues they found, how they measured the issue, what they did to address it, and what are the residual risks. Registering those factors not only allows an organization to demonstrate its due diligence, but also allows later scrutiny of its decisions.

When writing down that information, organizations might benefit from following best practices in software documentation. As the “Write the Docs” hub of software documentation recommends, the contents of good software documents should:

- **Avoid repeating information** that is available in other sources, such as the software code, unless some degree of repetition is beneficial for understanding.
- Keep in mind that **readers tend to skim the documentation** for useful examples and quick answers before reading it in depth.
- **Be consistent** with other sources in language and format.
- Be correct and reflect the current state of the software; **incorrect documentation is worse than nothing.**

Finally, those documents should be drafted in a way that allow their readers to find the information that is contained in them. Burying information amid the documentation runs against the spirit of those disclosure requirements and can easily become a resource drain for the supervisory authority and the supervised organization itself. As such, it should be avoided both for its practical wastefulness and for the risk of sanctions for non-compliance.

Session 11.2. Disclosure duties towards downstream developers

> By the end of this session, learners will be able to **outline** when and why the developers of AI models and systems are obliged to supply information to other actors who want to incorporate those products into their own systems.

Any AI system, no matter how small it is, is the product of a **complex value chain**. As we have seen in Part II of this training module, the creation and use of AI involves a variety of technical steps, and often relies on models and other components developed by third parties. This means that the actors at the end of this value chain do not always have visibility into the inner arrangements of the components they use. For example, if **InnovaHospital** decides to use a ready-made large language model to create a chatbot, the company supplying that model is unlikely to grant full access to the model’s configuration. Nonetheless, the hospital would still be responsible for the data processing it controls.

Data protection law and the AI Act both feature some mechanisms to address the potential information gaps ensuing from this situation. They do so by requiring that organizations supplying AI models and other components disclose some information about those components to the actors that incorporate them into their own systems. Data protection officers (DPOs) overseeing AI-driven initiatives should be aware of these legal requirements to safeguard user rights and meet regulatory standards.

Supply chain disclosure under the GDPR

Under the GDPR, developers of AI systems, when acting as data processors, must support downstream data controllers in fulfilling their obligations to respond to data subject rights requests. According to Article 28(3) GDPR, a data controller that hires a processor to carry out a task must lay down by contract (or other legal act) the conditions under which that processing will take place. This includes the need to adopt safeguards.

Consider a situation in which the university **UNw** decides to hire a contractor to develop a plagiarism detection system for its exams. Not only will the university retain its responsibilities as a data controller, but it will also need to specify safeguards that must be followed by the contractor. Those safeguards might include technical measures, such as those discussed in the next unit of this training module. But any controller would do well to require the processor to supply some information that might be essential for the controller's own compliance with legal requirements. They might also consider establishing protocols for communication between the organizations, to ensure smooth investigation of any future issues.

Likewise, the GDPR also requires an **explicit division of competences** in cases of joint controllership. Under Article 26(1) GDPR, joint controllers must clearly define their respective responsibilities for compliance. For instance, a healthcare AI model provider working jointly with a hospital to process patient data must determine who will be responsible for communicating the data collection and usage terms to patients, ensuring that both parties uphold the GDPR's transparency requirements. In this case, a controller might be able to avoid sharing information with its joint controllers. They cannot do so, however, at the expense of the information that must be supplied to data subjects and regulators.

Additional requirements under the AI Act

In the context of high-risk AI systems under the AI Act, further disclosure requirements apply. Article 25 of the AI Act stipulates that if a high-risk AI system's purpose is repurposed by a downstream provider, the original provider is partially relieved of compliance obligations. However, they must still cooperate by providing essential information about the AI system to help the new provider meet regulatory standards.

For example, if a financial institution repurposes a high-risk AI system initially developed for fraud detection to assess credit risk, the original developer must share information on the model's intended capabilities, limitations, and risks to support proper usage. Still, it is the financial institution that will be responsible for ensuring that the system complies with the applicable legal requirements when it is used for credit risk assessment.

For general-purpose AI models, Article 53 AI Act imposes a broader obligation to supply documentation and information. That documentation and information must be kept up-to-date and made available to providers who intend to use the model in their own systems. The bare minimum content of that disclosure is specified in Annex XII AI Act:

1. A general description of the general-purpose AI model including:

(a) the tasks that the model is intended to perform and the type and nature of AI systems into which it can be integrated;

(b) the acceptable use policies applicable;

(c) the date of release and methods of distribution;

(d) how the model interacts, or can be used to interact, with hardware or software that is not part of the model itself, where applicable;

(e) the versions of relevant software related to the use of the general-purpose AI model, where applicable;

(f) the architecture and number of parameters;

(g) the modality (e.g. text, image) and format of inputs and outputs;

(h) the licence for the model.

2. A description of the elements of the model and of the process for its development, including:

(a) the technical means (e.g. instructions for use, infrastructure, tools) required for the general-purpose AI model to be integrated into AI systems;

(b) the modality (e.g. text, image, etc.) and format of the inputs and outputs and their maximum size (e.g. context window length, etc.);

(c) information on the data used for training, testing and validation, where applicable, including the type and provenance of data and curation methodologies.

As we have seen in the previous units, a data controller will need this kind of information to carry out their various duties. Without the kind of information listed in Point 2 above, a controller will be unable to assess whether the use of that model poses specific risks in the intended context or supply meaningful information about the AI system. Therefore, this AI Act requirement supports compliance with data protection requirements, regardless of the risk level of the application in which the model is used.

Session 11.3. Technical disclosure and the right to an explanation

By the end of this session, learners will be able to **distinguish** between the various interpretations given to the "right to an explanation" in legal scholarship and practice.

One of the distinctive features of EU data protection law is that it grants a variety of rights to data subjects. To implement these rights, the GDPR creates a series of obligations for the data controllers who process data pertaining to those subjects. Those obligations remain valid when processing is done by an AI system, and they are supplemented by some additional requirements laid down in the AI Act. In this session, we will discuss what kinds of information data controllers must disclose to data subjects about their use of AI.

From a perspective of timing, it is important to distinguish between two moments of disclosure. Data controllers must disclose information about processing done by an AI system, whether the data has been collected from the data subject or not.[7] Additionally, data subjects have the right to request access to their personal data being processed and to information about processing.[8]

Because those rights are connected to actual processing operations, they must be exercised regarding the controller of that processing. That is, the data controller(s) for processing during the training stage will supply information about personal data used to train the AI system, while the data controller(s) of the deployed system will supply information about its use in each context.[9] In that regard, AI systems are treated just like any other form of processing.

What is unique about disclosure duties, when AI systems are involved, is the so-called "right to an explanation." Recital 71 GDPR mentions that data subjects should have, at least, the right to an explanation of automated decisions, but such a right does not appear in Article 22 GDPR. As a result, there was considerable controversy about whether such a right exists. That controversy is likely to be settled in definitive by the forthcoming decision of the European Court of Justice in Case C-203/22 (*Dun & Bradstreet Austria*), in which one of the referred questions deal precisely with the extent to the right to an explanation.

In the meantime, the dominant view among academics (see, e.g., Kaminski 2019) and data protection authorities (see Vale and Zanfir-Fortuna 2022) is that such a right can be grounded in the right to access to "meaningful information" about the logic of

[7] Articles 13 and 14 GDPR, respectively.
[8] Article 15 GDPR.
[9] On the identification of those actors, see Session 6.1 of this training module.

automated decision-making.[10] This right does not apply to all AI systems, but it applies "at least" in cases of automated decision-making under Article 22 GDPR, which are often carried out with AI. Therefore, at least some AI systems are subject to this rule.

The clause "at least in those cases" in Article 15(1)(h) GDPR suggests that a data controller might have an obligation to disclose "meaningful information about the logic involved", as well as "the significance and envisaged consequences of processing", even when processing does not qualify as automated decision-making. In a more restricted reading, one could understand this clause to merely state that data controllers *can* disclose that kind of information in other contexts. While this is certainly true, this possibility becomes an obligation in some cases.

Recently, the European Court of Justice broadened the understanding of "automated decision-making" under Article 22 GDPR. In the case C-634/21 (*Schufa*), it has ruled that a credit score calculated from personal data could be considered "automated individual decision-making" when a third-party receives that score and draws strongly on it to establish, implement, or terminate a contractual relationship. That is, an AI system (or any other form of data processing) that strongly influences a decision can be covered by Article 22 GDPR even if a human theoretically has a say in the process.

Additionally, Article 86 AI Act establishes that any affected person subject to a decision taken on the basis of the output of a high-risk AI system, which produces legal effects or similarly significantly affects that person, has the right to obtain "clear and meaningful explanations" of the role the AI system plays in the decision and the main elements of the decision taken. This right has been designed as a safeguard for cases that are not covered by the GDPR's right to an explanation,[11] and it requires a narrower form of disclosure. The deployer does not need to explain the logic guiding the decision, just the role of AI and the contents of the decision itself.

The concept of "meaningful information" about an AI system's decision logic

The determination of what counts as "meaningful information" under the GDPR is necessarily contextual. That is because access to that information is a data subject right, and as such it must be thought from the subject's perspective. The information provided about the decision logic must allow the subject to make sense of processing and how it affects their rights, liberties, and interests. It must give data subjects the grounding to decide whether to exercise their other rights, such as the right to contest an automated decision (Bayamlioglu 2022). If an explanation is to be successful in that aim, it must meet certain formal and substantive requirements.

[10] Article 15(1)(h) GDPR.
[11] Article 86(3) AI Act.

On the formal side of things, an explanation must be presented in a way that a data subject can understand. But data subjects can come from a variety of backgrounds. The average individual cannot be expected to have the time or the technical competences to understand technical explanations, so disclosing model parameters or a system's source code will not contribute to their understanding of the system. On the other hand, a technically savvy individual, or a person working with a civil society organization, might have the resources for a more in-depth exploration of technical issues. So, they will likely be unsatisfied with an explanation directed at laypersons.

Given this broad range of data subject capabilities, organizations would do well to follow a multi-layered approach to disclosure (Kaminski and Malgieri 2021). Doing so would entail preparing information that can be digested at various levels of complexity, and supplying that information according to data subject needs, on request. That will ensure that data subjects that need basic information are not smothered in technical detail, while other data subjects can dig deeper within their rights.

On the substantive side of things, the requirements are much less clear. The main question that is raised (see, e.g., Brkan and Bonnet 2020) is whether the disclosure of the "meaningful logic" behind an automated decision can happen without revealing the system's inner workings. On a literal reading of the requirement, that seems to be the case. An abstract description of how the system produces its inputs from its outputs might be enough to give an actionable view of why things have been decided in one way and not in another. However, compelling arguments have been made, both by academics and data protection authorities, that more disclosure is needed. Once again, the decision in C-203/22 (*Dun & Bradstreet Austria*) will provide more legal certainty in this regard.

Elements of meaningful information

In the absence of well-defined legal requirements in this regard, we will conclude this session by discussing some kinds of information that can be useful for compliance. Because both the technology and the legal elements of this issue are moving fast as of 2024, learners should make sure to check whether particular solutions are still applicable or if they have been replaced by something else. Still, it is possible to offer some considerations that will likely be relevant for any understanding of "meaningful information" under data protection law.

A primary component of disclosing "meaningful information" about an AI system is explaining the **inputs** that system considers as it produces its own outputs. For example, if an AI system assesses creditworthiness, it may consider inputs like income, credit history, and recent transactions. Communicating these inputs to data subjects provides them with a clearer understanding of the data influencing decisions about them, enabling them to verify the accuracy of their personal data and, if necessary,

request corrections. This transparency also helps ensure that data processing complies with principles of fairness, as individuals can better understand how relevant information impacts the outcomes they receive.

In addition to disclosing inputs, data controllers should communicate **how different inputs could lead to different outcomes**. While it is not always feasible to explain complex AI model logic in detail, providing examples or scenarios can help illustrate how certain changes in input data might affect the AI system's output. For instance, explaining that a credit assessment score could be different if income or employment status were updated gives individuals a practical sense of the decision-making logic, without delving into technical complexities. Such explanations are valuable, as they give individuals a tangible understanding of the system's logic, particularly in contexts where AI may influence significant aspects of their lives. Session 12.2 of this training module will discuss some technical measures that can support organizations in generating this kind of explanation.

It is also crucial to clarify **how the AI output impacts real-world decisions**. Data controllers should indicate whether the AI system's output is directly applied to make decisions or if it serves as a recommendation subject to human review. For example, an AI-driven hiring system may rank candidates based on qualifications, but a hiring manager makes the final selection. Distinguishing between direct and mediated applications of AI outputs helps individuals understand the role of human oversight in decision-making processes, fostering greater transparency about how their data is used.

While explainable AI methods and other technical means for transparency can assist in making these processes more transparent, they are not the sole solution. An explanation that covers these essential elements—inputs, potential outcome variability, and application context—is often sufficient to fulfil GDPR obligations without requiring deep technical detail. However, data controllers must balance this transparency with the need to protect trade secrets: their own secrets or those of the upstream providers from whom they acquire models and other components. Balancing the duty of disclosure with the need to respect those secrets can be a tricky challenge in practice. Still, it is a challenge organizations need to face in order to comply with data protection law when they use AI.

Conclusion to Unit 11

This Unit has covered many types of disclosure duties that are present in data protection law and the AI Act. The various forms of opacity discussed in Unit 4 of this training module all come into play here, creating obstacles to deployers and providers of AI technologies. It is in the best interest of those organizations to adopt measures that secure the information they need to disclosure. The bad news is that the disclosure

obligations remain in force even though AI makes things much more complicated. The good news is that there are various measures that can contribute to disclosure.

A few of those are relevant to many, if not all, of the modes of disclosure we have considered above:

- Maintaining **comprehensive and updated documentation** of processes and decisions.
- Keeping in mind **how the recipients of documents and other forms of disclosure will use the information** and preparing it accordingly.
- Using documents for the **clear definition of responsibilities** throughout the supply chain.
- Rely on **examples and context** to make information more accessible.
- Rely on a **multi-layered approach to disclosure**, in which the same information can be presented in ways that are more accessible to each kind of stakeholder.

Each of these practices has its own obstacles. For example, a multi-layered approach creates the challenge of ensuring that all forms of disclosure remain coherent with one another. Still, for the most part, the previous sessions illustrate how disclosure remains possible even in a world of opaque AI everywhere.

Prompt for reflection

UNw is considering incorporating a third-party AI model into its admissions process. However, it worries about its ability to ensure compliance with GDPR transparency requirements when it relies on an external provider. How should it manage its relationship with the third-party provider to ensure compliance with GDPR and AI Act requirements? What types of information should the third-party provider share with the university to enable transparency with students and regulators? Discuss the role of contracts, such as those mandated under Article 28(3) GDPR, in securing access to information in AI supply chains.

References

Article 29 WP, ‘Guidelines on Transparency under Regulation 2016/679’ (2018).

Emre Bayamlıoğlu, ‘The Right to Contest Automated Decisions under the General Data Protection Regulation: Beyond the so-Called “Right to Explanation”’ (2022) 16 Regulation & Governance 1058.

Andrew Bell and others, ‘It’s Just Not That Simple: An Empirical Study of the Accuracy-Explainability Trade-off in Machine Learning for Public Policy’ in (ACM 2022) FAccT ’22 248.

Adrien Bibal and others, ‘Legal Requirements on Explainability in Machine Learning’ (2021) 29 Artificial Intelligence and Law 149.

Sebastian Bordt and others, ‘Post-Hoc Explanations Fail to Achieve Their Purpose in Adversarial Contexts’ in (ACM 2022) FAccT ’22 891.

Maja Brkan and Grégory Bonnet, ‘Legal and Technical Feasibility of the GDPR’s Quest for Explanation of Algorithmic Decisions: Of Black Boxes, White Boxes and Fata Morganas’ (2020) 11 European Journal of Risk Regulation 18.

Melanie Fink and Michèle Finck, ‘Reasoned A(I)Administration: Explanation Requirements in EU Law and the Automation of Public Administration’ (2022) 47 European Law Review 376.

Margot E Kaminski, ‘The Right to Explanation, Explained’ (2019) 34 Berkeley Technology Law Journal 189.

Margot E Kaminski and Gianclaudio Malgieri, ‘Algorithmic Impact Assessments under the GDPR: Producing Multi-Layered Explanations’ (2021) 11 International Data Privacy Law 125.

Blazej Kuzniacki and others, ‘Towards eXplainable Artificial Intelligence (XAI) in Tax Law: The Need for a Minimum Legal Standard’ (2022) 14 World Tax Journal 573.

Gabriela Zanfir-Fortuna, ‘Article 13. Information to Be Provided Where Personal Data Are Collected from the Data Subject’ in Christopher Kuner and others (eds), *The EU General Data Protection Regulation (GDPR): A Commentary* (Oxford University Press 2020).

Gabriela Zanfir-Fortuna, ‘Article 14. Information to Be Provided Where Personal Data Have Not Been Obtained from the Data Subject’ in Christopher Kuner and others (eds), *The EU General Data Protection Regulation (GDPR): A Commentary* (Oxford University Press 2020).

Gabriela Zanfir-Fortuna, ‘Article 15. Right of Access by the Data Subject’ in Christopher Kuner and others (eds), *The EU General Data Protection Regulation (GDPR): A Commentary* (Oxford University Press 2020).

Unit 12. Regulating AI by Design

By the end of this unit, learners will be able to **compare** different approaches to data protection by design, **identify** the problems they are able to address, and **combine** them at various junctures of the life cycle of an AI system.

The EU's approach to the regulation of digital technologies gives considerable attention to how these technologies are designed. In the field of data protection, this attention manifests itself in two core legal requirements. Article 25(1) GDPR establishes a requirement of **data protection by design**, as it obliges data controllers to adopt technical and organizational measures that address risks to data protection. A similar logic can be seen in Article 32(1) GDPR, which creates a **security by design** obligation to adopt measures (both technical and organization) to avoid cybersecurity issues. Both sets of obligations are directed towards data controllers, who must identify the risks created (or amplified by processing) and choose the best measures to address them.

Data controllers do not make those choices in a vacuum. Both data protection by design and security by design establish that the controller must consider, such as the likelihood and severity of risks or the technological state of the art. Nonetheless, it places data controllers in the position of specifying how those legal provisions need to be interpreted in specific technical contexts.

Another flavour of regulation by design is present in the AI Act. Its rules on high-risk AI systems and general-purpose AI models with systemic risk both establish certain technical requirements that must be met before commercialization. The same is true of the supplementary rules that Article 50 AI Act creates for systems regardless of their risk classification. But, unlike the GDPR, the AI Act focuses on the adoption of **technical** measures in the AI system. The three approaches to regulation by design (security by design, data protection by design, and the AI Act's technical requirements) coexist, as they are all obligatory at the same time. This raises questions about whether and why these design mandates might clash with one another.

All forms of regulation by design used in EU law create ongoing obligations. Article 25 GDPR requires providers to adopt measures both at the moment of processing and when the means for processing are determined, while the AI Act obliges providers and deployers throughout the entire life cycle of an AI system. Both approaches to regulation by design also cover a broad range of values. The GDPR is designed to protect data subjects from the impact that processing might have on their fundamental rights, while the AI Act has the explicit aim of safeguarding health, safety, and public values such as the protection of fundamental rights, democracy, and the rule of law.

To cover all those values, regulated actors will need to use several types of technical and organizational measures for each context. For example, some systems might be able to benefit from anonymised or synthetic data, but a system that generates profiles will necessarily involve personal data. Even when personal data is intrinsic to the application, organizations can still adopt measures and safeguards to protect it to the greatest extent possible. For example, an organization developing an AI system would need to adopt cybersecurity measures to prevent leaks of personal data, while a deployer organization could adopt controls over who has access to AI outputs. Each application will be better served by a different mix of measures, but some best practices can be useful for a broad range of applications.

This unit introduces some examples of technical interventions that might promote data protection even if falling outside the scope of PETs. Some of these measures are oriented towards data subjects, allowing them to play a more active role in the defence of their rights:

- Explainable AI techniques (discussed in more detail in Session 12.2) can help data subjects in obtaining more information about automated decisions.
- User interfaces might be used to allow users to exercise some of their rights (such as access to information) with no intermediation.
- Online dispute resolution tools might be a path to allow individuals to exercise their right to contest automated decisions, especially in digital environments.

Other measures are directed, instead, at the needs of data controllers:

- Technical documentation can allow an organization to understand what is going on within a system.
- Some design choices, such as the use of inherently interpretable models, can facilitate compliance with information disclosure duty.
- Control over parameters of a system might be used to implement different balancing acts between values (for example, sacrificing some efficiency for extra accuracy).

This unit examines three sets of techniques that fall under the broad umbrella of data protection by design. **Session 12.1** discusses privacy-enhancing technologies (PETs), geared towards the minimization of personal data processing. **Session 12.2** deals with other technical and organizational approaches that foster aspects of data protection that go beyond privacy, such as the exercise of data subject rights. **Session 12.3** then wraps up the unit with an overview of technical and organizational approaches that aim to ensure fairness in AI systems.

Session 12.1. Privacy-Enhancing Technologies (PETs)

By the end of this session, learners will be able to **identify** different privacy-enhancing technologies (PETs), **illustrate** their contribution to data protection, and **recognize** their conceptual and technical limits.

Privacy-enhancing technologies (PETs) are technical methods developed to reduce the impact of data processing on individual privacy. In general lines, those methods foster privacy by minimizing the amount of data processed in each operation and ensuring the confidentiality and security of any data that is processed. This session will explore how the use of such technologies can contribute to data protection compliance when AI systems are designed and used.

Many PETs are used for broader purposes than the development of AI systems. For example, there are various techniques for data anonymization, which remove identifying factors in a way that prevents the data from being associated with a natural person. Differential privacy, on the other hand, adds "noise" to data queries, masking individual entries while preserving the overall utility of the data. These methods underscore the value of controlling data access as a means of reducing privacy risks.

Additionally, some AI-specific techniques have been designed with privacy in mind. For example, federated learning enables machine learning models to train across decentralized data sources without transferring data directly to a central system. This approach reduces data exposure while still allowing AI models to benefit from diverse datasets.[1]

Organizational measures as part of a privacy arrangement

While PETs are powerful tools, they are only one aspect of effective data protection. They must be paired with organizational measures that foster privacy and data security. Internal practices such as training personnel on the responsible handling of data, tracking data access, and setting restrictions on who can interact with AI models that process personal data are essential.

By training staff to handle data responsibly and implementing logging systems that track data access, organizations can create a culture of accountability that complements their technical measures. Additionally, controlling and monitoring access to AI systems helps prevent unauthorized data use and supports compliance with data protection regulations.

[1] Once again, learners interested in technical detail would do well to consult Enrico Glerean, *Elements of Secure AI Systems.*

However, certain privacy risks cannot be mitigated by organizational measures alone. For instance, data protection professionals should be aware of the European Data Protection Board's (EDPB) Recommendations 01/2020, which caution that measures like access controls are vulnerable to tampering by state actors or external adversaries. This vulnerability highlights the need to evaluate when privacy risks might require changing or limiting the use of certain AI technologies altogether.

Depending on the limitations of technical and organizational measures, an organization might need to consider whether it needs to abandon its (planned) use of AI. For example, it might be impossible to create a system that automatically allocates scholarships to students based on their academic performance, if that system cannot be designed in a way that does not discriminate between them in a way prohibited by law. In that case, the necessary design measure is not designing (or using) the AI system in the first place. Sometimes, the only winning move is not to play.

The limits of privacy-enhancing technologies in data protection

Despite their advantages, PETs have limitations that data protection professionals must carefully consider. Some privacy-preserving techniques are in **pilot stages of development** and are not ready for deployment in practice. For instance, although homomorphic encryption—allowing computations on encrypted data without exposing it—shows promise, it remains too complex and resource-intensive for widespread use. Until these emerging PETs become more practical, organizations may need to be cautious with them or be transparent about their limitations to ensure a realistic understanding of compliance. Other PETs, instead, are more mature and can be used more extensively.

An important conceptual limitation of PETs is their focus on data minimization, a key principle in privacy. **Minimizing data collection aligns well with privacy goals but does not capture the entire spectrum of GDPR obligations**. For example, some of the informational rights of data subjects discussed in Session 11.3 of this training module require providing information about how the system considers the circumstances of data subjects. To keep that information accessible, one needs to reduce the overall degree of confidentiality promoted by the system, creating a trade-off between privacy-as-confidentiality and data protection's goal of promoting control over the use of personal data (Veale et al. 2018). Suppressing data purely for the sake of minimization could inadvertently restrict individuals' rights and weaken the protection of fundamental rights overall. Thus, data protection officers need to weigh the benefits of minimization against the need to maintain a balanced approach to all GDPR principles.

Ultimately, while PETs do not fulfil every compliance need, they are valuable tools that can significantly reduce privacy risks. Informed use of PETs, combined with robust organizational measures and a clear understanding of their limits, allows data protection

officers to align AI systems with legal obligations. PETs should be seen not as standalone solutions but as part of a multi-faceted approach to comprehensive data protection in the AI era.

Session 12.2. Technical measures for AI transparency

> By the end of this session, learners will be able to **exemplify** techniques that promote technical transparency in AI systems and **assess** whether those techniques are adequate considering the relevant data protection risks.

In Unit 11 of this training module, we have seen that data protection law and the AI Act feature a broad range of information disclosure requirements. There is no one-size-fits-all solution, as data subjects, authorities, and society as a whole need several types of information, which they will use for different purposes. This unit examines whether and how design-based interventions can contribute to compliance with those information duties.

Technical interventions might be necessary to the extent that some of the information data controllers must provide refers to the inner workings of an AI system or model. As we discussed in Session 8.2 of the module, there is some controversy about the extent to which data controllers must provide detailed information about how the model operates, or if it is sufficient to provide highly abstract information. For some purposes, such as closer audits by data protection authorities, abstract information is not enough. Whenever that is the case, data controllers will need to deal with the technical opacity of AI.

One can distinguish between two sets of technical approaches that can be useful for this purpose. On the one hand, **explainable AI** (XAI) approaches try to distil the complexity of an AI system into key factors that determine its action. On the other hand, **interpretable AI** changes the system itself, building it with a simpler model that can be made legible to humans instead of a complex system based on more arcane machine learning techniques. Each of these approaches to the technical complexity of AI technologies has its pros and cons, which we will now consider.

Explainable artificial intelligence and the right to an explanation

XAI models offer a scientific approach to the black box problem. They start from the fact that we often do not know how complex AI systems work. Even if we set up their general architecture and training parameters, the sheer scale of those models, and the fact that they undergo a long training process, means that nobody—not even a trained expert—will have immediate access to everything that happens within an AI system. To solve these problems, XAI techniques aim to reconstruct the decision procedure and offer an understandable account of what is going on (Holzinger et al. 2022). If and when

they succeed, the ensuing model contributes to our understanding of the complex system that is being explained.

To achieve this goal, researchers have proposed a dizzying array of technologies. A review from a few years ago (Holzinger et al. 2022) identified at least seventeen methods that were in current use as of 2020. Some of these approaches are **model-agnostic**, that is, they try to reconstruct what a model does based on its outputs. For example, the LIME technique (Holzinger et al. 2022, p. 15) tries to represent the predictions of a complex model, such as a deep neural network, in terms of a surrogate model that is much simpler to understand than the original one. Anchor models try to identify "if-then" decision rules that capture the behaviour of a complex model. Those and other techniques end up creating surrogates that can be used for understanding the original AI model.

Other explanations are contingent on certain features of the models they explain. Layer-wise relevance propagation (LRP) approaches, for example, offers a procedure through which one can simplify the underlying logic of a larger model. To do so, it requires information about that model's internal arrangements (Holzinger et al. 2022, p. 18). The ensuing explanation is potentially more complex than what a model-agnostic explanation would offer, but the access to model-specific information allows the explanation to reflect more of the original model's actual functioning.

Model-agnostic and model-sensitive XAI techniques both advance scientific understanding of what is going on within AI models. This kind of understanding, however, is not necessarily equal to what the law demands when it establishes a "right to an explanation." Most XAI technique aim at a scientific explanation of the models they explain, that is, they supply potential mechanisms that would explain what the model does (Creel 2022). The legal conception of a right to an explanation is, instead, related to the *justification* of a decision: whether it is compatible with legal requirements.[2] There are some reasons to believe one kind of insight does not always lead to the other.

Some recent works (in particular, Bordt et al. 2022) have suggested that **XAI methods cannot be trusted in adversarial contexts**. In such contexts, the data subject's interest in discovering how an AI system works is contrary to the data controller's interest in preserving that information. For example, **InnovaHospital** might want to prevent a patient from understanding an AI diagnosis tool for several reasons, such as avoiding a lawsuit from a misdiagnosis or protecting intellectual property. Whenever that is the case, the data controller has various possibilities for manipulating the outputs of the explanation model. The use of XAI would not be enough to ensure trust and would

[2] See Session 8.2 of this course.

need to be accompanied by technical and organizational measures to reduce the controller's possibilities of manipulation.

Another problem is that **XAI techniques are not necessarily more understandable than opaque models**. A study on the legibility of AI models (Bell et al. 2022) has found that many people find that "simpler" models are still too complex to understand. As such, they are not necessarily more accessible or insightful than the bigger models they aim to replace. The use of XAI technologies must therefore make sure that any outputs are understandable for the audiences they are meant to reach. If that is not possible, then the use of XAI might not be an answer to the legal demands for explanation.

Inherently interpretable models

If XAI techniques are not enough to provide transparency, what can be done? One approach to that problem is the use of **inherently interpretable models**. Even though many advanced AI applications are powered by complex, opaque AI models, there are many important problems that do not require all that complexity. In fact, computer scientists such as Cynthia Rudin (2019) have shown that, for some tasks, simpler models can perform at least as well as black box models. Whenever that is the case, data controllers have fewer reasons to rely on the opaque alternatives, especially for sensitive tasks.

The move towards simpler models can be desirable for several reasons, such as reduced costs in development and execution. However, **its usefulness for transparency will depend on the audience** to which information is meant. The same concerns with legibility discussed above were also identified when users were exposed to interpretable AI models (Bell et al. 2022, Kolkman 2022). Still, these models might be more legible than black box alternatives for technical experts, who have the technical baggage needed to make sense of them. They might also be useful for investigative journalists, who can experiment with the parameters of AI models and find out how they operate. Therefore, reliance on inherently interpretable models can be beneficial even if it those models are not necessarily more accessible for laypeople (Busuioc et al. 2023).

Session 12.3. Designing for algorithmic fairness

> By the end of this session, learners will be able to **exemplify** technical approaches that can be used for the design of fairer algorithms.

This session examines some design measures for addressing fairness issues in AI models. As we have examined in Session 10.3 of the module, algorithmic fairness is a complicated problem, both for its technical challenges and for the difficulty in representing legal understandings of fairness in a way that can be measured and

implemented in an AI system. Nonetheless, some technical approaches can promote fairness, or at least mitigate known risks such as algorithmic discrimination and biases.

Best practices in addressing risks to fairness (such as Snoek and Barberá 2024) emphasize the need to address issues throughout the entire life cycle of an AI system. That is, responses to fairness issues are not restricted to the development process or to the initial deployment. Hence, one must look at all the life cycle stages examined in Part II of this training module.

Fairness interventions at the inception stage

At the inception stage, fairness can be pursued in many ways. Organizations can evaluate whether the purposes they pursue with an AI system or model are not, in themselves, discriminatory. For example, an AI system that is designed to carry out an unlawful form of discrimination cannot be salvaged by any technical measures.

Organizations might also want to examine how they frame the problem(s) that they want AI to solve, in order to avoid abstraction traps (Snoek and Barberá 2024, p. 20), that is, situations in which the design ignores important aspects of reality. For example, if **InnovaHospital** wants to create an AI system to assess heart attack risks, it needs to take into account the differences in symptoms between men and women.

Fairness in design and development

When it comes to the development of an AI system, fairness practices can be directed towards the data used in training, the development of the algorithmic system, and the documentation of system design decisions. All of those are useful not just for avoiding potential sources of unfairness in algorithmic predictions, but also to keep track of design decisions that are relevant for accountability and for future updates to ensure the system continues fair.

The foundation of a fair AI system lies in the quality of its data:

- Ensuring **completeness** is essential, as gaps in data can lead to skewed or biased outputs. For example, a university admissions model at the **UNw** university might underperform for certain demographic groups if its training data lacks sufficient examples of applicants from those groups.
- Similarly, **accuracy** in labelling and data collection is crucial to avoid embedding errors into the system.
- **Representativeness** is another key aspect: datasets should reflect the diversity of the real-world populations the AI system will serve. For instance, **DigiToys** must ensure its AI-driven toys are tested on a diverse range of children to avoid unintended exclusion or stereotyping.

The virtues of good documentation we discussed in Session 10.1 of the module are also relevant for promoting fairness. All relevant decisions and assumptions made during the AI lifecycle should be recorded systematically. Such **comprehensiveness** ensures transparency and enables future audits. The language used in documentation also matters: it should be **accessible** to diverse stakeholders, avoiding overly technical jargon while ensuring clarity. Furthermore, keeping documentation **up to date** is essential, as decisions about data, algorithms, and design choices must be revisited in response to evolving societal and regulatory contexts. For example, **InnovaHospital** might track updates in medical guidelines or regulatory changes to ensure its diagnostic models remain fair and compliant.

Finally, fairness during the design of an AI system requires attention to the model training process. Designing fair algorithms involves employing ***appropriate fairness metrics*** to measure and address potential biases. Despite the issues discussed in Session 10.3 of the training module, quantitative and qualitative metrics (such as those proposed by Wachter et al. 2021, Weerts et al. 2023) can be helpful to diagnose certain issues with algorithmic decisions.

Understanding the **sources of bias** is equally important. Research on algorithmic biases has proposed various forms in which the training processes, and the decisions that guide them, can skew the operation of an AI model. For example, a learning bias happens when an AI model prioritizes some metric over other objectives that the system must pursue, such as prioritizing effectiveness over fairness (Snoek and Barberá 2024, p. 17). Those biases can be amplified later, as humans have their own cognitive biases. For example, people overseeing AI systems often override decisions that “look wrong” while deferring to algorithmic decisions that conform to their biases (Alon-Barkat and Busuioc 2023). Fair development of AI systems will therefore require attention both to technical biases and those affecting human-computer interaction.

Fairness during and after the initial deployment

Once an AI system has been deployed, its core design is generally fixed. At this stage, promoting fairness involves ensuring that the system's outputs are applied in ways that align with equitable outcomes. However, one must still pay attention to potential fairness issues related to the AI outputs themselves. This is the case for two reasons:

1. Some sources of unfairness might have **escaped detection** during the development process. If they go unchecked in deployment, they might only be noticed after they have harmed data subjects.
2. Even if a system were perfectly fair at first, **unfairness might appear after deployment**. This might happen as part of a model’s self-learning processes because the data that was originally relevant no longer is so, because society has changed, or many other factors.

As such, organizations need to keep an ongoing surveillance of whether their AI systems and models are processing data fairly.

During the deployment process, organizations can promote fairness by **testing their new systems in real-world conditions**. By doing so, they can verify whether it functions as intended across diverse settings. For example, **DigiToys** might evaluate how its interactive toys perform in households with varying languages and cultural norms, ensuring consistent and appropriate interactions. Similarly, **InnovaHospital** could test its diagnostic models across diverse patient demographics to confirm equitable performance. If any issues are detected, further work might be needed on the system. Alternatively, an organization might adjust its procedures to avoid unfairness, for example by improving human oversight once the system is deployed.

After the system is deployed, an organization needs to evaluate what biases might emerge during operation. For instance, an admissions algorithm at **UNw** could inadvertently reinforce pre-existing inequalities in access to education if societal biases are reflected in the input data or institutional policies. **Regular assessments** help identify and address such contextual biases.

To ensure continued fairness, organizations should continue **to track fairness metrics after deployment**. For example, as societal norms or data patterns evolve, an AI system might need recalibration to avoid perpetuating outdated or unfair assumptions. Bias detection should be an ongoing effort, incorporated into the risk management practices discussed in Unit 8 of this training module.

Finally, **ongoing interaction with regulators and affected communities** is vital *to* maintaining fairness and accountability. Engaging directly with those impacted by the AI system helps organizations understand real-world fairness concerns and adapt to shifting regulatory and societal expectations. For example, **DigiToys** could collaborate with parents' groups to address concerns about how its AI systems influence children's behaviour, while **InnovaHospital** might consult healthcare regulators and patient advocacy groups to align its practices with ethical standards. Relying on those actors will help any organization in finding fairness issues that escaped its own monitoring tools.

Conclusion to Unit 12

"Regulation by design" is a concept that is in vogue nowadays, and not without reason. If problems can be solved by technical means, this contributes to a higher level of protection for data subjects, reducing possibilities of human error and subversion. Not all problems can be solved by technical measures, for a series of reasons, such as the limits of what one can represent in a computer language, the current state of the art, and potential conflicts between the values that the various by-design approaches are

meant to protect. Nonetheless, technical design remains a powerful tool for compliance, as shown by the various interventions discussed in this unit.

To recapitulate the key points of our discussion:

- Privacy-enhancing technologies (PETs) seek to maintain the utility of a system while reducing the amount of personal data it processes.
    - Some technologies, such as differential privacy approaches to the training set, homomorphic encryption, and federated learning, might be particularly relevant in the context of AI.
    - The development of PETs is grounded on a view of privacy as concealment, which is not entirely aligned with the idea of control in data protection. Hence, the use of PETs is not enough to discharge all data protection obligations.
    - However, they contribute, at very least, to data minimization, and so an organization might want to consider the extensive use of PETs where it makes sense.
- Technical interventions cannot fully remove opacity, but they can reduce the efforts needed to understand the inner workings of AI systems.
    - Explainable AI (XAI) models try to offer a scientific explanation of the main factors behind an AI approach. Various techniques have been proposed, but they struggle to deal with adversarial contexts that are common in precisely the kind of situation that is likely to create legal issues.
    - Inherently interpretable models are obtained by building systems without the use of black-box techniques. This is not always possible, given the success of black-box models such as neural networks for some problems. Yet, there are many applications in which those opaque models do not necessarily perform better than the alternatives.
    - In addition, some empirical research suggests that neither approach is really intelligible to the general public. Technical transparency might nonetheless be beneficial for technical experts, as well as for actors such as courts, supervisory authorities, and investigative journalists.
- Despite the various challenges to algorithmic fairness, some multidisciplinary teams have developed approaches that are feasible in practice and address some real unfairness concerns.
    - This research is often grounded on US law, and as such it is not always directly applicable for compliance with EU law.
    - There are extensive mappings of biases that can emerge during the design process, and addressing those biases can contribute to fairness.
    - Any approach to fairness in AI will require constant recalibration as technologies and social expectations change.

It might be the case that the different values promoted by each by-design approach clash with one another. The conflict might be conceptual: for example, privacy by design might require the elimination of some information that might be useful for the exercise of other rights (Veale et al. 2018). But it might also emerge because of limited time and resources that do not allow designers to meet all needs equally. Solving these conflicts will require careful consideration, which also requires engagement with branches of the law beyond data protection. Still, to find the ideal equilibrium, one must consider the entire life cycle of the AI system rather than look just at immediate needs. Otherwise, today's solution will likely become tomorrow's compliance problem.

Prompt for reflection

Reflect on a real-world scenario (or one of the hypothetic cases of **UNw**, **DigiToys**, or **InnovaHospital**) where implementing regulation by design principles might lead to a conflict between privacy, transparency, and fairness. How should organizations prioritize these principles in their AI systems? What strategies can be employed to mitigate the risks associated with favouring one principle over another? Are there any contexts where one principle might justifiably take precedence?

References

Marco Almada and others, 'Art. 25. Data Protection by Design and by Default' in Indra Spiecker gen. Döhmann and others (eds), *General Data Protection Regulation: Article-by-article commentary* (Beck; Nomos; Hart Publishing 2023).

Saar Alon-Barkat and Madalina Busuioc, 'Human-AI Interactions in Public Sector Decision-Making: "Automation Bias" and "Selective Adherence" to Algorithmic Advice' (2023) 33 Journal of Public Administration Research and Theory 153.

Andrew Bell and others, 'It's Just Not That Simple: An Empirical Study of the Accuracy-Explainability Trade-off in Machine Learning for Public Policy', *FAccT '22* (ACM 2022).

Sebastian Bordt and others, 'Post-Hoc Explanations Fail to Achieve Their Purpose in Adversarial Contexts', *FAccT '22* (ACM 2022).

Lee A Bygrave, 'Article 25. Data Protection by Design and by Default' in Christopher Kuner and others (eds), *The EU General Data Protection Regulation (GDPR): A Commentary* (Oxford University Press 2020).

Alessandra Calvi and others, 'The Unfair Side of Privacy Enhancing Technologies: Addressing the Trade-Offs between PETs and Fairness', *FAccT 2024* (ACM 2024).

Luca Deck and others, 'A Critical Survey on Fairness Benefits of Explainable AI', *FAccT 2024* (ACM 2024).

Pierre Dewitte, ‘The Many Shades of Impact Assessments: An Analysis of Data Protection by Design in the Case Law of National Supervisory Authorities’ (2024) 2024 Technology and Regulation 209.

Ernestine Dickhaut and others, ‘Lawfulness by Design – Development and Evaluation of Lawful Design Patterns to Consider Legal Requirements’ [2023] European Journal of Information Systems Early Access.

EDPB, Guidelines 4/2019 on Article 25 on Data Protection by Design and by Default (2020).

ENISA, Best Practices for Cyber Crisis Management (2024).

Daan Kolkman, ‘The (in)Credibility of Algorithmic Models to Non-Experts’ (2022) 25 Information, Communication & Society 93.

Efstratios Koulierakis, ‘Certification as Guidance for Data Protection by Design’ (2024) 38 International Review of Law, Computers & Technology 245.

Andreas Holzinger and others, ‘Explainable AI Methods - A Brief Overview’ in Andreas Holzinger and others (eds), *xxAI - Beyond Explainable AI* (Springer 2022).

Christina Michelakaki and Sebastião Barros Vale, ‘Unlocking Data Protection By Design & By Default: Lessons from the Enforcement of Article 25 GDPR’ (Future of Privacy Forum May 2023).

Cecilia Panigutti and others, ‘The Role of Explainable AI in the Context of the AI Act’, *2023 ACM Conference on Fairness, Accountability, and Transparency* (ACM 2023).

Cynthia Rudin, ‘Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead’ (2019) 1 Nature Machine Intelligence.

Suzanne Snoek and Isabel Barberá, ‘From Inception to Retirement: Addressing Bias Throughout the Lifecycle of AI Systems. A Practical Guide’ (2024).

Michael Veale and others, ‘When Data Protection by Design and Data Subject Rights Clash’ (2018) 8 International Data Privacy Law 105.

Sandra Wachter et al., ‘Bias Preservation in Machine Learning: The Legality of Fairness Metrics under EU Non-Discrimination Law’, 123 W. VA. L. REV. 735, 744 (2021).

Unit 13. Data Protection and Large Language Models

By the end of this unit, learners will be able to:

- **distinguish** large language models from other kinds of AI systems.
- **identify** current open problems of those models from a data protection perspective; and
- **exemplify** potential solutions to addressable issues.

In the last few years, much of the discussion about AI has focused on Large Language Models (LLMs). An LLM is a type of AI model that is designed to process and generate text, using vast amounts of training data to learn how to extract and reproduce patterns in human language. Those models were brought to public attention in November 2022, when OpenAI released the first iteration of ChatGPT, a chatbot powered by one of its earlier LLMs. Because ChatGPT could answer a variety of queries and dialogue with its users, it soon became a popular tool, used not just for fun but incorporated into a series of applications.

As these models have grown larger and more complex, their development has concentrated within a few major corporations due to the significant computational resources and expertise required. Many businesses now use these models through APIs or tools without needing to develop them in-house. This concentration raises unique challenges for data protection, as responsibilities and risks are distributed across the supply chain in ways that can complicate compliance with data protection law.[1]

LLMs raise a variety of challenges to data protection law. Of those, a few are particularly critical.

- A large language model might **expose sensitive data** used during their training process, for example if an attacker crafts a prompt that gets the model to output that data.[2]
- The operation of an LLM might **violate data protection principles**. One example can be seen in the so-called hallucinations, that is, on false outputs generated out of the blue by such models. These false results compromise the accuracy principle, especially when they sound plausible to the observer.

[1] In addition to risks that go beyond individual privacy and data protection, such as their potential for misuse and the creation of systemic risks. For instance, LLMs can be used to generate realistic but entirely false content, fuelling misinformation or political disinformation campaigns. These capabilities can undermine public trust, incite political instability, and even manipulate individuals' opinions by spreading fake news or impersonating public figures. Such risks are covered by the AI Act's rules on general-purpose AI models, as well as by other legal instruments, but they exceed the data protection focus of this training module.
[2] See Kucharavy et al. (2024, ch. 7).

- LLMs might also create **obstacles to compliance** with relevant legal requirements. For example, an organization that uses an LLM that is opaque to it will struggle to comply with the various forms of disclosure covered in Unit 12 of this training module.

Concentration of the markets for LLM technologies can compound those risks, as it means that an organization's compliance challenges will be affected by the technical decisions made by a handful of technology suppliers. As such, these suppliers become especially relevant for data protection enforcement. This does not mean, however, that the organizations using those LLMs suddenly become exempt from their GDPR duties. Just that compliance with those duties might become more difficult.

In addressing those issues, organizations and regulators face specific challenges. One of them is the difficulty in ascribing responsibility. Providers of LLMs have the technical expertise to modify them but often lack the specific context of each business or end-use case. Without knowing the particularities of each deployment, it is difficult for model providers to foresee every privacy risk or ensure compliance with regulations.

Meanwhile, organizations using LLMs within their own systems—whether as part of their operations or as end-users—understand the context of their data processing and how the model affects their workflows. What they typically lack is the ability to alter the model itself to address context-specific privacy risks. For example, a company using a commercial LLM for customer service could struggle to adjust the model's data retention practices to meet GDPR requirements, as the model's provider controls these technical aspects.

An emerging alternative to the LLMs provided by large technology companies is the rise of open-source LLMs, which organizations can modify more freely than commercial models. Although open-source LLMs may not perform as consistently as closed-source models from major tech firms, they offer growing capabilities and flexibility. With open-source LLMs, companies have more control over model adjustments, which can aid in adapting data handling to meet legal requirements.

Open-source LLMs bring their own challenges. Developers and businesses using them will need a greater degree of technical expertise, not just to make use of these models but also to ensure their compliance with technical requirements. For example, a model provider's measures for complying with data protection requirements will propagate to any AI systems using those models. An organization that uses its own models will need to implement their own measures for that purpose, and to do so they might need to make changes to the original model. If, on the one hand, open-source models give them the power to make such changes, on the other hand that power is of little use if the organization lacks the expertise to do so. Each organization must therefore consider

what kind of model, if any, is better suited for data protection compliance considering the resources it has available.

In this unit, we examine the data protection issues created by LLMs and how they affect the design and use of AI systems based on them. In **Session 13.1**, we consider the opportunities and risks created by those models, obtaining a clearer view of what they can and cannot do with the current state of the art. **Session 13.2** then looks at data protection issues that emerge during the development of LLMs. To wrap up the unit, **Session 13.3** discusses measures that organizations can adopt when they use systems based on LLMs.

Session 13.1. The opportunities and risks of large language models

> By the end of this session, learners will be able to **exemplify** why the use of LLMs might be desirable in some cases and why it might not be in others.

Large language models (LLMs), such as OpenAI's GPT-4, represent a significant advancement in artificial intelligence. Those models are produced by training deep neural networks on large datasets of written text, which often include almost everything that is publicly available on the internet for a given language.[3] This training process allows LLMs to recognize patterns in text data, which they subsequently apply to the consumption and generation of other texts. Current state-of-the-art LLMs can perform an impressive range of language tasks, including text summarization, answering questions, translating languages, generating written content, and even conducting conversations that mimic human dialogue.

These capabilities make LLMs particularly valuable in areas where efficient language generation is required. For example, **InnovaHospital** might want to create a chatbot that interacts with patients to carry out an initial triage based on their symptoms, while **DigiToys** might benefit from LLMs to enable their toys to better interact with children. Those tasks can be useful and even transformative of certain kinds of work. However, they create a series of risks for data protection, too.

Despite their impressive capabilities, LLMs are not omniscient or capable of reasoning like a human. They do not inherently understand concepts or the world in the way humans do; rather, they generate responses based on probability distributions derived from their training data. As a result, they lack "real" intelligence or understanding and are prone to "hallucinations" — confidently generating incorrect or fabricated information. This is not merely a flaw but a fundamental limitation that cannot currently be eliminated in their design.

[3] See Kucharavy et al. (2024, ch. 1).

As a result, these models remain limited in their reliability and interpretative abilities. LLMs can sometimes provide misleading or inaccurate information. For example, even the best current LLMs often struggle with simple mathematical problems, depending on how they are phrased. Not to mention the countless examples of “hallucinations” one can find online. Reliance on misleading information generated by an LLM runs against the GDPR principle of accuracy in the processing of personal data, especially if that wrongful output becomes the basis for subsequent processing.

Another data protection concern with LLMs relates to how those models are trained. There is, currently, a scientific controversy on whether a large language model does, in fact, memorize information during its training process. If that is the case, then the model itself might contain personal data. This is because LLMs are trained on massive datasets, which often contain public and potentially private information about information. As a result, those models might capture personal data used in its training, even if precautions are taken in the process. However, the technical and legal question of whether memorization matters for data protection law remains open.

Even if no memorization takes place, LLMs can themselves be taken as a source of personal data. If an LLM can generate outputs that refer to an identified or identifiable natural person, that output might qualify as personal data. This is the case even if the output itself is false. After all, the definition of personal data refers to information that can be associated with a natural person, and accuracy is a requirement that applies *after* something is classified as personal data. The outputs of an LLM, or of a system that is based on an LLM, might therefore be subject to data protection requirements. Currently, there is intense discussion among EU data protection authorities on how to apply data protection requirements in this context, and further guidance is likely to come soon.

Because they can be used for a broad variety of purposes, LLMs are likely to qualify as general-purpose models under the AI Act.[4] And, given the potential reach of those systems that can generate language at scale, they might create systemic risks. The AI Act’s regulatory approach to this kind of model is pretty narrow. It establishes a set of security and safety requirements,[5] which must be observed by the most powerful models in the market.[6] This approach ensures that the state-of-the-art models are made safe(r) before they can be commercialized or made accessible to the public.

The AI Act’s response to systemic risk goes, in some respects, beyond the individual focus of data protection law. Even so, the brunt of the legal response to the risks

[4] See the definition in Article 3(63) AI Act.
[5] Article 55 AI Act.
[6] As they depend on quantitative metrics that are currently met by a handful of models, such as GPT-4o or Google’s Gemini.

created by LLMs falls on the metaphorical shoulders of data protection law. This is because many of the potential harms from the training and use of LLMs can affect identifiable persons. In the following sessions, we consider how LLMs affect compliance duties for providers and deployers of AI systems based on LLMs.

Session 13.2. Safeguarding measures during model development

> By the end of this session, learners will be able to **describe** some of the data protection issues that might emerge during the development of an LLM and **sketch** potential responses to them.

The training process of LLMs raises some important data protection questions. As discussed in Session 6.1 of this training module, the LLM's provider is likely to qualify as the data controller of any personal data processed during the training procedure. Consequently, they need to adopt safeguards to address general risks involved in model training,[7] as well as the risks that are specific to LLMs.

The first obligation of a data controller training an LLM is to ensure that any personal data in the training set can be lawfully processed. This entails identifying whether such data is present in the training set, finding proper legal bases for its processing, and ensuring the observance of the applicable rights and requirements from data protection law. In this regard, things are similar to any other AI model or system. What changes is that some specific factors become more salient.

One of them concerns the data sources used to train the LLM. Language models are often created from "data trawling" (Kuru 2024), that is, the mass consumption of publicly available data. While public data may seem readily available for training, it does not automatically mean it can be freely used; public availability does not negate data protection obligations. Data controllers should assess each data source, verifying that **proper legal grounds**, such as legitimate interest or consent, justify the use of personal data within the dataset. This step is especially important if the data includes sensitive or special categories of personal data. Documenting the legal basis for data use can support compliance and prepare organizations for potential audits or inquiries.

Data minimization is another critical principle in the development of LLMs. This means using only the minimum amount of data necessary to achieve the training objectives. Privacy-enhancing techniques, such as those discussed in Session 12.1 of this training module, can be used to reduce the amount of personal data contained in the training dataset while still retaining the dataset's utility. Organizations should consider implementing robust data governance frameworks that regularly evaluate and update

[7] See Session 6.3 of this training module.

data minimization practices throughout the LLM's development and maintenance phases.

LLMs are often designed for **general-purpose applications**, meaning they can be adapted to a wide range of uses beyond those initially anticipated. This versatility raises unique data protection concerns because a model's potential misuse or unintended applications could lead to new privacy risks. Under the AI Act, such general-purpose AI systems are subject to specific regulatory requirements, which mostly relate to the kind of information disclosure discussed in Unit 11 of this training module. For the most advanced AI models, some additional requirements apply, as stipulated in Article 55 AI Act.

Regardless of the applicability of those provisions, developer organizations must anticipate a range of potential use cases for their model, including harmful applications. Given that the developer is likely to be a controller for processing in the training process, the purposes for which a model is tailored are relevant for determining the lawfulness of processing. Additionally, these developers will also need to consider kinds of misuse that, while not intended, are reasonably foreseeable from the original application. For example, **DigiToys** needs to consider that any safety issues with their model might allow hackers to interact with the children that play with their smart toys.

In forecasting potential misuses or errors, organizations should pay particular attention to possible **unintended processing of sensitive data**. Without adequate controls, a model could inadvertently generate or expose sensitive information such as health, racial, or political data about individuals. This matters even if the information generated by the system is not true, as in that case the LLM will be generating (and potentially help spread) false information about an individual. For example, if an LLM is used to generate texts that slander someone, the generated outputs are personal data even if they have no basis on reality, as they are associated with an identifiable natural person. As such, developers might want to consider the use of techniques to restrict the outputs, prohibiting certain kinds of prompts or interactions with the model.

Moreover, **models accessible to minors** might pose unique risks, as they could unintentionally generate harmful or age-inappropriate content. Effective measures to address these risks include implementing monitoring mechanisms that detect and flag potential data-related issues in real-time. Additionally, integrating ethical and responsible use guidelines into the deployment phase can help prevent misuse, ensuring the model's output aligns with data protection standards and user safety expectations.

To meet both GDPR and AI Act obligations, it is critical to adopt **specific safeguards**. For instance, transparency measures can help users understand the limitations and potential biases within the model. Furthermore, to fulfil GDPR joint controllership

obligations, organizations must collaborate with partners to jointly establish accountability measures for data processing.

Organizations may also consider adopting **robust model documentation**, as discussed in Session 10.1 of the module. Those documents can detail the training data, potential risks, and safeguards to enhance the model's transparency. Although this step is more exhaustive than documentation requirements for traditional data processing activities, it is crucial to clarify accountability and manage systemic risks associated with general-purpose AI models under data protection law the AI Act.

An emerging issue in LLM development is the risk of **poisoning the well**, that is, the degradation of the data used to train a model. Degradation can happen, for instance, when AI models are trained on data that includes substantial amounts of AI-generated content. Some recent studies have suggested that models trained on AI-generated data can collapse in performance over time. But this phenomenon can also happen deliberately, as seen in Session 3.3 of this training module. An attacker might poison the data used to train a model in order to alter and manipulate a model’s operation, for example to make it generate content that discriminates against a specific ethnic group. In both cases, the changes to an AI model’s training set can impact its performance and/or create undesirable side effects.

The issue of poisoning the well suggests organizations need to exercise caution when using synthetic data to train an LLM, adopting data auditing practices and tracking metrics to ensure that the model outputs achieve a sufficient level of quality. An alternative approach is to design and use datasets that are strictly curated and verified to contain only authentic human-generated content, reducing the risk of model degradation.

For **open-source LLMs**, data protection responsibilities remain relevant and, in many cases, complex. While identifying the controller or processor roles can be challenging in collaborative open-source settings, these models are still subject to GDPR obligations, as Article 2(2) GDPR excludes very few scenarios. This means that even when a model is freely accessible and collaboratively developed, accountability measures must still be enforced.

Under the AI Act, open-source systems benefit from certain exemptions, such as reduced documentation requirements. These exemptions apply only to the finished systems themselves, and LLMs remain covered by the rules applicable to general-purpose AI models. Some of the documentation requirements in Article 53 AI Act do not apply to open-source models, but others do. Furthermore, all rules on systemic risk apply to models that meet the relevant thresholds. Therefore, open-source projects aimed at developing LLMs, especially at the state of the art, need to pay attention to data protection and AI Act requirements.

Session 13.3. Safeguarding measures for model use

By the end of this session, learners will be able to **outline** risks that must be assessed when an organization considers whether to deploy an LLM.

As discussed throughout the training module, even the best design practices cannot eliminate all data protection risks created by AI. This means that organizations deploying systems based on LLMs, or incorporating LLMs into the AI systems they develop, must themselves adopt some safeguards. To do so, they must understand the roles the LLM plays within the data processing architecture. In this session, we will discuss some measures that can be useful for data protection professionals as they seek the best responses to the risks created in specific contexts.

When an LLM is integrated as part of software development, particularly during fine-tuning, privacy concerns surrounding the data used and generated by the model are paramount. Organizations should understand how the LLM processes, stores, or shares data and where it fits within the broader workflows that use it. The first step towards such an understanding is a thorough evaluation of the **instructions for use** that the LLM developer is required to supply to downstream providers.[8] Based on the issues flagged by the developer, an organization can consider initial safeguards to adopt and identify issues that require additional investigation.

To supplement that documentation, an organization might want to conduct its own **black box tests and audits** of the models they intend to do so. While those kinds of tests do not provide the same levels of insight that white-box practices can supply,[9] they can help organizations flag some issues before they lead to harms in practice. Such a mapping exercise helps identify potential points of data exposure, especially if personal data is involved.

When **fine-tuning** a model, an organization must evaluate whether the data it uses for that purpose contains personal data. If so, it will need to determine the proper basis for processing data for a purpose distinct than the one that motivated the original processing of that personal data.[10] For example, if a customer support system uses a fine-tuned LLM to provide responses based on prior interactions, the organization will need to obtain the consent of those data subjects or establish the presence of a suitable basis for further processing, as required by Article 6(4) GDPR.

A critical component of deploying LLMs responsibly involves **scrutinizing the input data** that will be fed into the model. Since LLMs often rely on vast amounts of data to

[8] See Session 11.2 of this training module.
[9] See Session 7.3 of this training module.
[10] See Session 6.2 of this training module.

improve responses or recommendations, it is essential to ensure that this data is free from unnecessary personal information. Implementing strict input filtering and anonymization protocols can reduce the risk of processing personal data inadvertently. For instance, sensitive identifiers such as names, addresses, and financial details should be either stripped out or obfuscated before data is inputted into the LLM. Additionally, data minimization principles must be observed, ensuring that only the essential data required for the model's functionality is used.

It is also important to evaluate **how the LLM provider might use any personal data** submitted to or generated by the model. Many LLM providers retain certain types of data to improve model performance or conduct diagnostics, which may lead to secondary processing risks. Organizations should scrutinize service-level agreements and data processing agreements to understand what, if any, access the provider has to this data and whether additional safeguards, like encryption and access controls, are in place.

To mitigate the risk of unauthorized data access, organizations may opt for **self-hosted models or models that can run locally**, ensuring that data does not leave their controlled environment. Furthermore, if the LLM provider will engage in processing the data, specific contractual clauses should be enforced to limit such access strictly to what is necessary for service delivery.

When using an LLM-powered tool, attention should also be given to the **data generated by the model itself**. The outputs generated by LLMs can potentially contain or create personal data, which poses additional data protection challenges. For instance, if an LLM generates summaries or recommendations that incorporate individual user preferences or behaviours, these outputs may qualify as personal data under data protection law if they can be associated with an identified or identifiable natural person. In this regard, it is essential to apply accuracy checks on generated content to prevent misinformation or inaccurate profiles. Regular monitoring of the model's outputs, coupled with ongoing adjustments to the model's parameters, can help maintain the reliability and relevance of the information produced.

Depending on the purposes of the system that uses the model, the organization might want to adopt **input filters** that prevent the model from being used for certain purposes such as the generation of hate speech. It might also adopt **output filters** that prevent some of the outputs generated by the LLM from reaching its end user, for example by preventing a chatbot from outputting swear words or discriminatory terms. However, one must also be aware that such methods are often subject to subversion ("jailbreaking"), which might or not be addressable through further design of the system's inputs.

To ensure compliance, organizations should also incorporate a process for **regular audits and reviews** of the LLM's performance and its handling of data. This includes both technical assessments, such as testing the effectiveness of data anonymization measures, and compliance reviews, such as verifying that user consent is up to date and aligned with the intended data processing purposes. Establishing a clear auditing trail for data inputs, outputs, and consent records will help maintain transparency and accountability in line with data protection laws.

Conclusion to Unit 13

In this unit, we have looked more closely at a specific type of AI model, which powers a growing number of applications that process personal data. The three sessions above are not enough to exhaust the complexities of the topic, but they provide a starting point for further analysis. By drawing on the discussions above, you will be able to flag issues that require further attention and investigate them in the context of particular systems. Furthermore, the same steps of analysis can be used to analyse other AI technologies that might be relevant to your work.

We can synthetize the main points of the previous discussion as follows:

- LLMs are trained from vast amounts of data.
    - Most often, that data is scraped from the internet and other public sources, but one must keep in mind that publicity does not exclude the need for lawfulness in data processing.
    - This information is likely to include personal data of individuals, and some of it might not even contain falsehoods about the identified (or identifiable) persons.
- LLMs operate as "black boxes," making it difficult to explain how they process data and generate outputs.
- The complexity of LLMs may hinder the exercise of rights such as access, rectification, and erasure.
- Many of the measures for risk mitigation and elimination discussed in previous units can be tailored to deal with the specifics of LLMs.
- The complexity of LLMs and their need for data has some implications for their diffusion.
    - Few actors might have the capabilities to develop or to host LLMs at scale.
    - Fine-tuning those models is considerably less expensive and is more accessible for organizations and individuals.
    - There is an asymmetry of knowledge between upstream and downstream providers, and so obligations need to be well-distributed between them.

- A downstream provider using a ready-made system powered by an LLM, or incorporating an LLM into their own system, must evaluate which safeguards and protective measures they can implement.
    - Those measures might include organizational measures regarding how the LLM is used.
    - They might also include technical measures such as filtering input and output data or fine-tuning the model to address known risks.
- Given the capabilities of advanced LLMs and their widespread use, some of them might trigger the AI Act's thresholds for systemic risk, triggering additional legal requirements for their providers.

To translate these insights into action, focus on building robust collaboration frameworks with AI developers. By-design measures such as those discussed in Unit 12 of this training module can be useful both for model developers and for system developers relying on an LLM provided by somebody else. And, fundamentally, compliance checks must cover every stage of the AI lifecycle. Finally, the fast evolution of these technologies means that more concrete guidance is likely to emerge in the near future. Therefore, it is particularly important to stay updated of recent developments in data protection law.

Prompt for reflection

LLMs create not only privacy-related risks but also systemic risks under the AI Act, such as the spread of misinformation or harmful applications at scale. How should organizations anticipate and address potential misuse of LLMs? Are technical safeguards, like input filtering, sufficient, or do they require broader societal and regulatory interventions? How can organizations mitigate these risks while still leveraging the advantages of LLMs?

References

CNIL's Q&A on the Use of Generative AI (18 July 2024). Accessed 26 September 2024.

Data Protection Commission. AI, Large Language Models and Data Protection (18 July 2024). Accessed 26 September 2024.

Mindy Nunez Duffourc, Sara Gerke and Konrad Kollnig, 'Privacy of Personal Data in the Generative AI Data Lifecycle' (2024) 13 NYU Journal of Intellectual Property & Entertainment Law 219.

EDPS, 'Generative AI and the EUDPR. First EDPS Orientations for Ensuring Data Protection Compliance When Using Generative AI Systems.' (European Data Protection Supervisor 3 June 2024).

Lilian Edwards and others, ‘Private Ordering and Generative AI: What Can We Learn From Model Terms and Conditions?’ (CREATe Working Paper, CREATe 2024).

Florence G’sell, ‘An Overview of the European Union Framework Governing Generative AI Models and Systems’ (Stanford Cyber Policy Center working paper, 20 May 2024).

Philipp Hacker and others, ‘Regulating Gatekeeper Artificial Intelligence and Data: Transparency, Access and Fairness under the Digital Markets Act, the General Data Protection Regulation and Beyond’ (2024) 15 European Journal of Risk Regulation 49.

Andrei Kucharavy and others (eds), *Large Language Models in Cybersecurity: Threats, Exposure and Mitigation* (Springer 2024).

Taner Kuru, ‘Lawfulness of the mass processing of publicly accessible online data to train large language models’ (2024) International Data Privacy Law.

Paul Ohm, ‘Focusing on Fine-Tuning: Understanding the Four Pathways for Shaping Generative AI’ (21 June 2024).

OWASP, Top 10 for Large Language Model Applications (OWASP 2025).

David Rosenthal, ‘Part 19: Language models with and without personal data’ (Vischer, 2024).

Ilia Shumailov and others, ‘AI Models Collapse When Trained on Recursively Generated Data’ (2024) 631 Nature 755.

Unit 14. Supporting the Lawful Use of AI

By the end of this unit, learners will be able to **explain** why non-binding sources such as technical standards, third-party certification, and codes of practice are relevant for compliance with AI-related legal requirements.

Furthermore, they will be able to **assess** whether and how conformity with such a scheme is suitable to organizational needs.

This session wraps up the training module by introducing learners to the various technical instruments that can support their management of AI-related issues. It does not focus on specific standards (especially as the harmonized European standards still have not been published) but discusses their legal value and limitations.

Both the GDPR and the AI Act are designed as **technology-neutral laws**. That is, their obligations are meant to cover present and future technologies within the scope of those laws. Their legal requirements are formulated in general terms, and the application of those terms to specific technologies is left to a later stage. As seen in Unit 13, this often means that data controllers are actively involved in this process of translating legal requirements into technical ones. But they are not the only actors involved in this process.

To comply with their legal obligations, data controllers can rely on a variety of secondary sources. They can consult academic works on data protection law, hire external consultants, among other possibilities. None of those is mandatory, but they all might supply gaps in the interpretation of the law that an organization can have.

Under EU law, some of those sources are given a privileged status. As **Session 14.1** examines, the AI Act stipulates that conformity with harmonized technical standards generates a presumption that the AI system or model complies with the legal obligations covered by the applied standards. An organization can still decide to depart from such a standard, but they will need to show their approach meets the essential elements laid down in the Act. So, there are advantages to following this source even if it is not mandatory.

This significant role of standards is unique to the AI Act, but the Act and the GDPR also grant a special status to other sources. **Session 14.2** examines the differences in the legal value that each of those instruments grants to certification procedures and self-regulation mechanisms. Then, **Session 14.3** wraps up the training module by introducing the learner to measures that the AI Act introduces to support innovation in AI technologies.

Session 14.1. Technical standards

By the end of this session, learners will be able to **explain** the legal value of various kinds of technical standards under European Union law. Equipped with those distinctions, learners will be able to **evaluate** what kinds of standards, if any, are suitable for their compliance needs.

Technical standards are documents that specify a way of doing something. For example, the European standard EN 124:2015 governs the production of manholes: it distinguishes between different classes of weight loads to which manholes might be subject and defines some attributes that manholes for each class might have. While a manhole does not have much in common with AI technologies, the latter can also be standardized to some extent.

A technical standard for AI would define the properties that a certain AI-powered technology must meet. For example, a standard for facial recognition technologies might stipulate the need to adopt techniques that detect bias, as well as accuracy levels that must be met for a successful application. As such, those standards can help organizations identify which metrics are relevant for their technologies, and what values these metrics should have.

Types of technical standards

A technical standard is a **written document** that lays down norms. This document must be written by *someone*, and it will have a specific format. Given the technical nature of a standard, understanding its contents is something that requires some specialized knowledge in the domain of application. For example, the target audience of AI technical standards is that of AI experts that will work in the development of AI audiences. Beyond this core of meaning, however, technical standards can take a variety of forms, depending on their audience. We will now discuss some types of standards that can be relevant for AI.

The first distinction between standards concerns the object being standardized:

- **Product standards**, such as the EN 124 discussed above, establish technical norms for a physical or digital object.
- **Process standards**, instead, deal with how an organization does things. For example, the famous ISO 9000 series of technical standards establishes various norms that organizations can follow to increase the quality of the products and services they offer. As such, they are meant to change practices within an organization.
- Another role that standards can play is that of establishing **shared concepts** and vocabulary within a technical field. For example, the standard ISO/IEC 15408-

1:2022 defines concepts and principles that should guide the evaluation of IT security.

These three examples illustrate that compliance to standards might require changes to the behaviour of different people within an organization.

Another relevant distinction concerns the goals laid down by a standard. **Design standards** are technical standards that specify how a particular goal must be achieved. For example, the TCP protocol stipulates how a computer must format data and the procedure it must follow to transmit data to another computer. If a computer tries to communicate through this protocol without following those specifications, it will not succeed.

This approach can be contrasted with that of a **performance standard**, which merely specifies goals but leaves the manufacturer free to choose how these goals will be met. For example, a pollution standard for cars might specify that a car cannot emit more than a certain volume of CO2 per kilometre. If the car emits less than that, it meets the standard regardless of the technologies used to move it.

A final distinction that is relevant for our purposes concerns the source of a given technical standard. Some standards are produced by private organizations or consortia of private organizations: for example, the Blu-ray standard was created by a group of companies led by Sony. Others are produced in a more collaborative way, by organizations formed by representatives of private (and sometimes, public) entities, such as the International Organization for Standardization (ISO) or the Institute of Electrical and Electronics Engineers (ISO). Finally, some standards are produced by public bodies, such as the US National Institute of Standards and Technology (NIST) or the national harmonization bodies in the European Union. The pedigree of a technical standard might have implications for its legal implications.

The legal value of standards

Under the GDPR, technical standards are not a particularly salient factor. Article 43 GDPR mentions that the Commission may adopt implementing acts laying down technical standards for certification mechanisms and data protection seals and marks. However, it does not directly establish the power to adopt technical standards of general value. Data protection authorities can compel data controllers and processors to adopt certain measures,[1] and they can adopt and authorize contractual clauses that might have technical implications.[2] Still, these powers do not include neither the elaboration of binding technical standards nor the power to oblige *all* data controllers to follow a certain standard. As such, the use of technical standards is rarely mandatory for

[1] Article 58(2) GDPR.
[2] Article 58(3)(g–h) GDPR.

compliance with the GDPR, even if organizations are still free to rely on those standards to help them interpret the technical implications of data protection.

The AI Act, instead, gives considerable value to a specific type of technical standards. Article 40 AI Act establishes that conformity with harmonized technical standards generate the **presumption of conformity** with the provisions of the AI Act covered by that technical standard. That is, a high-risk AI system or a general-purpose AI model that follows the applicable standards is assumed to comply with the AI Act unless it can be shown otherwise. As one can expect, following such standards is therefore a way to reduce the effort involved in understanding what the AI Act requires of a data controller or processor.

This presumption only applies to **harmonized technical standards** (or parts of them) that have their references published in the Official Journal of the European Union. That is, an actor that follows a private standard such as ISO 42001 must still demonstrate that the measures they took meet the requirements of the AI Act. The standards that trigger the presumption are only those published by two European Standardization Organizations—CEN and CENELEC—in response to a request by the European Commission. These standards are unlikely to be made public before the end of 2025.

Additionally, the European Commission has the power to emit common specifications. In terms of content, a common specification is just like a technical standard—and as such, an organization is free to decide whether to follow it or not. What distinguishes it is the form of its adoption. The Commission can only create a common specification if it finds a technical issue that is not adequately covered by the harmonized technical standards it requested, and it must follow a specific legal procedure. But, if and once such a specification is adopted, it also generates a presumption of conformity with the legal requirements covered by it.

Under the AI Act, organizations developing or deploying AI technologies have strong incentives to follow harmonized standards or common specifications when they exist. Why might they rely on other kinds of standards, then? A few reasons might explain that:

1. **Following an international standard might be desirable or needed to reach markets beyond the EU***.* It might even be required by the laws of some other country in which an organization operates or sells its products. For example, China has its own approach to regulation of AI technologies.
2. **The EU standards might not be detailed enough**. In this case, the non-harmonized standard can help an organization make sense of the information it needs to implement the harmonized standard.

3. **An obligation might not be covered by a harmonized standard**. Given that harmonized standards for AI are shaped by the AI Act, they might not cover all the data protection risks examined above.

These conditions suggest organizations have good reasons to rely on sources beyond the forthcoming harmonized standards. When they do so, however, they must take care to show that the measures they adopt are enough to comply with the relevant legal requirements. Furthermore, regardless of the reliance on standards, they remain obliged to comply with data protection law. Following a standard (harmonized or not) does not eliminate this need but can be a useful tool in demonstrating compliance.

Session 14.2. Other mechanisms to support compliance

> By the end of this session, learners will be able to **identify** when third-party certification of an AI system is required under EU law. They will also be able to **describe** the key features of self-regulation mechanisms.

When seeking information about their legal obligations, organizations can rely in sources beyond technical standards. In this session, we will discuss two of the sources. First, we will consider how **certification schemes** can help actors in demonstrating their compliance and in obtaining information about the content of their obligations. Then, we will discuss how documents such as **codes of practice** and **codes of conduct** can guide organizations as they interpret their duties. Reliance on certifications and documents is not mandatory, but it is sometimes encouraged by legal advantages. As such, it is important to know how those arrangements work and whether they are suitable for a particular context.

Certification schemes as a tool for demonstrating compliance

Broadly speaking, certification is a process in which an organization relies on **a trusted process** to evaluate a product or a service that the organization offers. It is an established practice in modern lives: the food we eat, the electronic devices we buy, and so many other things often have certificates meant to reassure us of their quality. The situation is not different when it comes to the digital world, as sellers might want to use certification to build trust in an innovative technology.

In data protection law, the primary role of certification is as a form of demonstrating compliance. Article 42 GDPR clarifies that the certification process is voluntary. Controllers and processors can choose whatever certification they want (or none at all). However, certifications issued by bodies compliant with the requirements laid down in Article 43 GDPR are considered when an organization must show that it observed the data protection requirements for a given processing. This allows Member States to

ensure a certain degree of quality for high-end certifications, while still leaving market actors free to pursue other arrangements.

The AI Act also falls short of making third-party certification mandatory. In fact, for most AI systems, conformity with the applicable legal requirements must be demonstrated by the providers themselves through an internal assessment procedure. Article 43 AI Act further clarifies that this procedure does not leave room for the involvement of a certification body. Instead, those bodies (called “notified bodies” under the AI Act) are only involved in specific cases. In particular, third-party assessment remains s mandatory if the product in which AI is used would be subject to such an assessment under other provisions of EU law. Whenever that is the case, such certification must be pursued before the system can be placed on the market, put into service, or used in the EU.

In the absence of such a mandate, third-party certification offers little advantage from the legal perspective of the AI Act. It might nonetheless remain desirable for social reasons. Subjecting your product to the scrutiny of a trusted third party might be a way to create trust in it. For example, **DigiToys** might want to undergo external certification in order to show to prospective buyers that its smart toys are safe enough to be used with children. Additionally, an organization might use third-party certification to double-check or supplement its own internal controls. At the end of the day, external certification is no substitute for internal due diligence but can be a powerful supplement to it.

Codes of practice and other self-regulation instruments

As the previous units of this training module have shown, the EU approach to AI regulation allows considerable flexibility for developers and deployers of AI technologies. For the most part, those actors are the ones who identify the relevant risks and how they are best addressed by technical and organizational measures. Regulators have the power to address situations in which those measures are insufficient to protect rights, freedoms, and interests affected by the use of AI. Still, to a considerable extent, this regulatory power is expected to help the regulated actors in finding the best way for compliance.

To that effect, Article 40 GDPR establishes that the EU Member States and their data protection authorities, the EDPB, and the Commission must encourage the drawing up of codes of conduct. These codes of conduct are to be drawn up by associations and other bodies representing categories of controllers or processors and offer guidance for the problems faced by that category. For example, a code of conduct regarding the processing of medical data would be useful for **InnovaHospital** as it deals with the design of its AI systems that process personal data. It might offer guidance about how to pseudonymize data, how to configure parameters to ensure appropriate levels of

accuracy, and so on. A code of conduct therefore offers a bridge between the general provisions of the law and the specifics of particular applications of data processing.

A code of conduct is a **voluntary commitment**. An organization, be it public or private, can choose whether it will follow a code drawn up by representatives of a category. For example, the GDPR does not oblige the university **UNw** to follow a code of conduct elaborated by the National Association of Universities.[3] However, data protection authorities play a supervisory role in the process of drawing up these codes. This means data protection authorities can offer guidance regarding the sector-specific issues that associations can identify and ensure the quality of a code of conduct. By following a code of conduct approved by a data protection authority under the GDPR's procedure, an organization can be sure that its processes reflect the best practices in data protection.

This volunteer approach to data protection means that authorities can tap into the technical and practical knowledge of domain experts, while still guiding them on data protection. As such, the overall level of data protection might benefit from the competition between different codes of conduct, as well as from the experiences of different sectors.

Such an approach has been extended by the AI Act, which features two types of codes. The first type of code is the **code of practice for general-purpose AI models**. According to Article 56 AI Act, the EU AI Office[4] must encourage and facilitate the creation of those codes of practice. They are meant to guide the proper application of the Act's provisions on general-purpose AI systems, detailing obligations laid down in Articles 53 and 55 of the Act.[5] Their elaboration involves the providers of general-purpose models, national authorities, civil society organizations, academics, and other interested parties.

Drawing on those perspectives, the resulting codes are expected to offer detailed instructions on what providers of general-purpose AI models must do to comply with the AI Act. They are expected to define specific objectives and measures that must be adopted.[6] They must also include specific metrics for tracking conformity to those objectives,[7] and the actors who embrace the code of practice must provide regular reports on how they implemented their commitments.[8] This means an organization that

[3] The organization might be obliged by other sources, for example if the association makes adhesion to the code a requirement for participation in it.
[4] An organ within the European Commission.
[5] See Session 11.2 of this training module.
[6] Article 56(4) AI Act.
[7] Article 56(4) AI Act.
[8] Article 56(5) AI Act.

adheres to a code of practice obliges itself to implement their AI models in a way that reflects current best practices on software development.

Compliance with a code of practice remains voluntary. Yet, embracing such a code can bring advantages to providers of general-purpose AI models. Until harmonized standards on general-purpose AI models are published, a provider can use a code of practice to demonstrate compliance with obligations. That is, the fact that an organization joined a code of practice and is up to date with its commitments will be enough to establish that it has fulfilled the obligations covered by those practices.

Once a harmonized standard is released under the procedure covered in Session 14.1 of this training module, the codes of practice lose this additional value. Even then, a provider will likely be compliant with relevant AI Act provisions if it follows an up-to-date code of practice. What changes is that the provider will need to show the concrete measures it has taken. Mere adhesion to the code of practice will no longer be considered enough.

Finally, the AI Act also allows organizations to adopt **codes of conduct**. Through these codes of conduct, organizations are expected to voluntarily pledge to follow some (or all) the requirements for high-risk AI systems, even if their system is not classified as such. Because the requirements are not legally binding on them, following a code of conduct does not generate a presumption of quality. Still, it is seen as a way to push organizations towards best practices against AI risks. This is why the AI Office and the Member States are expected to encourage and facilitate the elaboration of such codes.[9]

Session 14.3. Measures supporting innovation in AI

> By the end of this session, learners will be able to **identify** potential sources for guidance as they deploy AI systems.

Part of the difficulty in regulating AI technologies comes from their novelty. Because AI allows for the automation of tasks that were previously outside the reach of computing, sometimes it can be difficult to figure out what can go wrong with a particular AI system or technology. Even when the risks are known, there is also uncertainty about whether the proposed fixes are sufficient to address them. After all, a solution that works well in a controlled test environment might not work so well in the real world. This creates a knowledge problem, which regulators try and address in a few ways.

Within the framework of the GDPR, data protection authorities have been active in providing guidance about factors relevant for the use of AI. The European Data Protection Board has edited various guidelines about legal requirements such as

[9] Article 95 AI Act.

automated decision-making, data protection by design and by default, and the legitimate interest basis for automated decision-making. Likewise, national authorities have often published guides about specific technologies, such as the generative AI systems discussed in Unit 13. A compliance plan for AI systems should refer to those documents whenever available.

Regulatory sandboxes in the AI Act

The AI Act includes some additional mechanisms for supporting organizations that intend to deploy AI systems. The first one is that of **regulatory sandboxes**, established in Article 57 AI Act. A sandbox is a controlled environment that can be used to assess emerging technologies before they are placed on the market or put into service. In this environment, an organization can experiment with the AI system in conditions resembling the real world. They will need to follow the testing protocols defined by the authority establishing that sandbox, which are meant to identify and fix risks before the system is put into widespread use. In this process, providers of AI systems are supported by the national regulatory authorities, which will offer guidance in identifying risks from AI and in complying with the applicable legal requirements.

Joining a sandbox can be advantageous for an organization that wants to develop an AI system or model. The first advantage is that a regulatory sandbox creates a **space for dialogue**: organizations can benefit from the expertise of regulators on the technical and legal issues raised by AI, and potentially benefit from the experiences of other organizations within the sandbox. In particular, the authorities responsible for a sandbox are required to help organizations diagnose potential risks to fundamental rights, health, and safety stemming from the use of AI.[10]

Within the sandbox, all legal requirements remain applicable. Organizations are expected to comply not only with the AI Act's requirements, but with the GDPR and any sector-specific legislation that covers their AI system or model. **Competent authorities still retain their supervisory and corrective power**s. However, they are expected to use their discretionary powers in a way that supports innovation.[11] Furthermore, the authorities involved in the sandbox cannot apply administrative fines to organizations that follow in good faith the sandbox's testing protocols.[12] **Regulators are therefore expected to guide organizations towards compliance**.

By 2 August 2026, each Member State of the EU is required to set up at least one sandbox for AI systems.[13] That sandbox can be a general sandbox for all kinds of innovative AI systems, but Member States can also set up separate sandbox for

[10] Article 57(6) AI Act.
[11] Article 57(11) AI Act.
[12] Article 57(12) AI Act.
[13] Article 57(1) AI Act.

different domains. For example, a state might create a sandbox for stimulating AI innovation in the medical sector and another one, following different rules, for innovations in education. The conditions that an organization must meet to enter the sandbox and to exit it (that is, to adopt a product that is cleared for use) are defined by the competent authorities for AI regulation. This means that such sandboxes might be extended to systems beyond the AI Act's high-risk classification.

The possibility of having sandboxes beyond high-risk systems is particularly useful because the sandboxes are not limited to AI Act enforcement. Under Article 57 AI Act, **the data protection authority must be involved in any sandboxes concerning personal data**. Likewise, sector-specific regulators must be involved in the sandboxes relating to their sectors of competence. Therefore, joining a sandbox allows organizations to understand what is required of them before they commercialize or put into service an AI system.

Processing personal data within sandboxes

Another major advantage of joining a regulatory sandbox is that it allows for the **further use of personal data**. Within a sandbox, a provider can lawfully reuse personal data collected for other purposes, if the AI system is meant to safeguard a substantial public interest in one of the following areas:[14]

> *(i) public safety and public health, including disease detection, diagnosis prevention, control and treatment and improvement of health care systems;*
>
> *(ii) a high level of protection and improvement of the quality of the environment, protection of biodiversity, protection against pollution, green transition measures, climate change mitigation and adaptation measures;*
>
> *(iii) energy sustainability;*
>
> *(iv) safety and resilience of transport systems and mobility, critical infrastructure and networks;*
>
> *(v) efficiency and quality of public administration and public services;*

This use of personal data is limited. The organization that wants to develop an AI system for those purposes within a sandbox must show that the use of such data is needed to meet the requirements for high-risk AI systems,[15] in particular by showing that alternative sources such as anonymized or synthetic data would be inadequate. It

[14] Article 59(1)(a) AI Act.
[15] Article 59(1)(b) AI Act.

must also adopt a series of safeguards for data use[16] and follow the testing protocols in the sandbox.

Real-world testing of AI systems

In addition to sandboxes, the AI Act also features a mechanism for testing AI systems in real-world conditions. Such tests are subject to a strict discipline, laid down in Article 60(4) AI Act. Those conditions include the need for approval of a testing plan (and, in many cases, registration) before any test can start, restrictions on the transfer of data to outside the EU, a limited duration for the test (at most six months, which can be extended by up to another six months upon a justified request), and safeguards for the testing subjects. Those subjects must provide their informed consent to participation in any such test.[17]

Market surveillance authorities are granted powers to **supervise the tests** and interrupt them if necessary. However, unlike the sandbox procedure stipulated above, real-world tests outside a sandbox are not necessarily integrated with data protection enforcement. As such, data protection requirements can be enforced normally, without the restrictions placed by sandboxes. Therefore, an organization might consider moving to this kind of testing only after it has established a solid basis for its processing of personal data.

Conclusion to Unit 14

The final unit of this training module has covered some tools and mechanisms that AI providers and deployers can use for making sense of the legal requirements in the GDPR and the AI Act. Because these legal instruments are designed to cover all sorts of circumstances, they cannot offer detailed guidance about every use case or technology. To supply this type of guidance, the legal instruments create some incentives that support private and quasi-private actors, such as standardization bodies, in providing tailored guidance. Relying on these sources is, of course, no substitute for organizational diligence, but they can be incredibly helpful for organizations as they try to comply with legal demands.

When it comes to technical standards, a few distinctions become relevant. The first distinction is between the harmonized standards that will be issued by CEN and CENELEC—which create a presumption of compliance with the relevant AI Act provisions—and other technical standards, which do not create the presumption of compliance but can be used for demonstrating that an organization followed best practices. It is also important to distinguish between standards that govern processes and standards that govern products, as well as between standards that lay down requirements and standards that lay down performance goals.

[16] Article 59(1)(c–j) AI Act.
[17] Article 61 AI Act.

Certification and self-governance mechanisms, such as codes of practice, can also be a source of guidance. They are not always granted the privileged status given to harmonized standards under the AI Act,[18] but they still contribute to compliance. These documents can distil the best practices available in industry and explain how they apply to specific contexts, helping regulated actors with interpretation. They can also be used as means to demonstrate the practices that an organization followed in design.

Finally, measures supporting innovation in AI—such as regulatory sandboxes, real-world testing, and facilitated compliance for SMEs—can reduce the legal barriers for the use of AI technologies. They allow organizations to benefit from guidance by regulators, while allowing regulators to learn more about risks before technologies are put into place. As such, joining them might be interesting, especially for organizations developing or deploying unproven AI technologies.

Ultimately, the decision on whether to rely on one or more of those tools falls to the organization itself. There might be good reasons not to pursue them, such as the cost of purchasing technical standards or pursuing extensive certification. Still, given the uncertainties surrounding AI technologies, they offer potentially valuable options for supporting any organization in its path to data protection compliance when using AI.

Prompt for reflection

Reflect on the differences between harmonized standards, international standards, and codes of conduct in the context of AI compliance. How might an organization like **DigiToys** decide which to adopt, and what factors should influence this decision?

References

Marco Almada and Nicolas Petit, 'The EU AI Act: Between the Rock of Product Safety and the Hard Place of Fundamental Rights' (2025) 62 Common Market Law Review.

Marta Cantero Gamito and Christopher T Marsden, 'Artificial Intelligence Co-Regulation? The Role of Standards in the EU AI Act' (2024) 32 International Journal of Law and Information Technology eaae011.

Peter Cihon and others, 'AI Certification: Advancing Ethical Practice by Reducing Information Asymmetries' (2021) 2 IEEE Transactions on Technology and Society 200.

Mélanie Gornet and Winston Maxwell, 'The European Approach to Regulating AI through Technical Standards' (2024) 13 Internet Policy Review.

Eric Lachaud, 'What GDPR Tells about Certification' (2020) 38 Computer Law & Security Review 105457.

[18] See, however, the temporary presumption created by the codes of practice for general-purpose AI models until the harmonized standards are published.

Sybe de Vries and others, ‘Internal Market 3.0: The Old “New Approach” for Harmonising AI Regulation’ (2023) 8 European Papers - A Journal on Law and Integration 583.