ML6

Navigating high-risk AI systems under the European AI Act: a guide for early stages

How to classify a high-risk AI system and how to deal with the requirements

In this white-paper, we explore how to deal with high-risk AI systems within the regulation of the European AI Act, ensuring compliance and fostering ethical AI innovation.

If you're immersed in the world of AI, the recent developments around the European AI Act are likely on your radar. If not, it's the right time to proactively think about the impact of the AI Act on your (future) AI project. The AI Act, adopted after extensive discussions among the EU institutions, marks a significant move towards regulating AI technologies across the European Union to balance citizen protection and technological advancement.

Which systems are high-risk AI systems in the light of the AI Act? And, why are these AI systems risk? How to deal with the requirements?

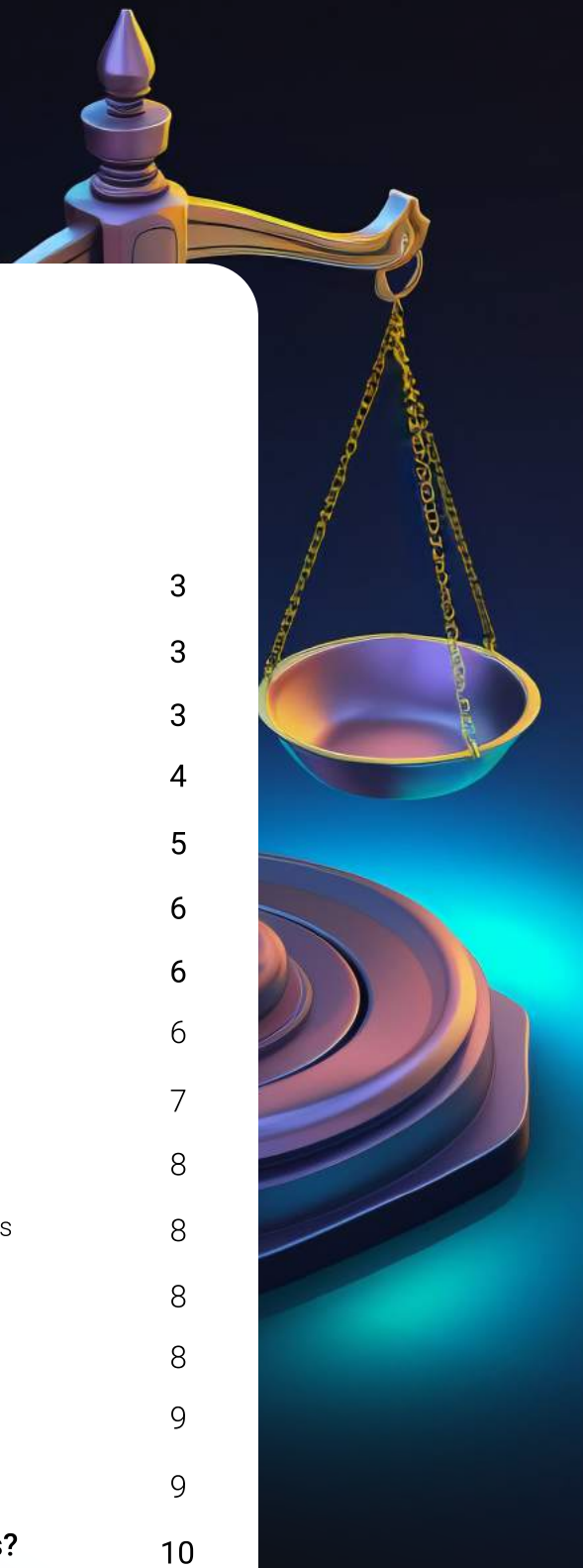


Table of Contents

What's the AI Act about?

If you're new to the AI Act, let's review the key points you need to understand. If you are already the AI Act expert of your company, feel free to skip ahead to page 4.

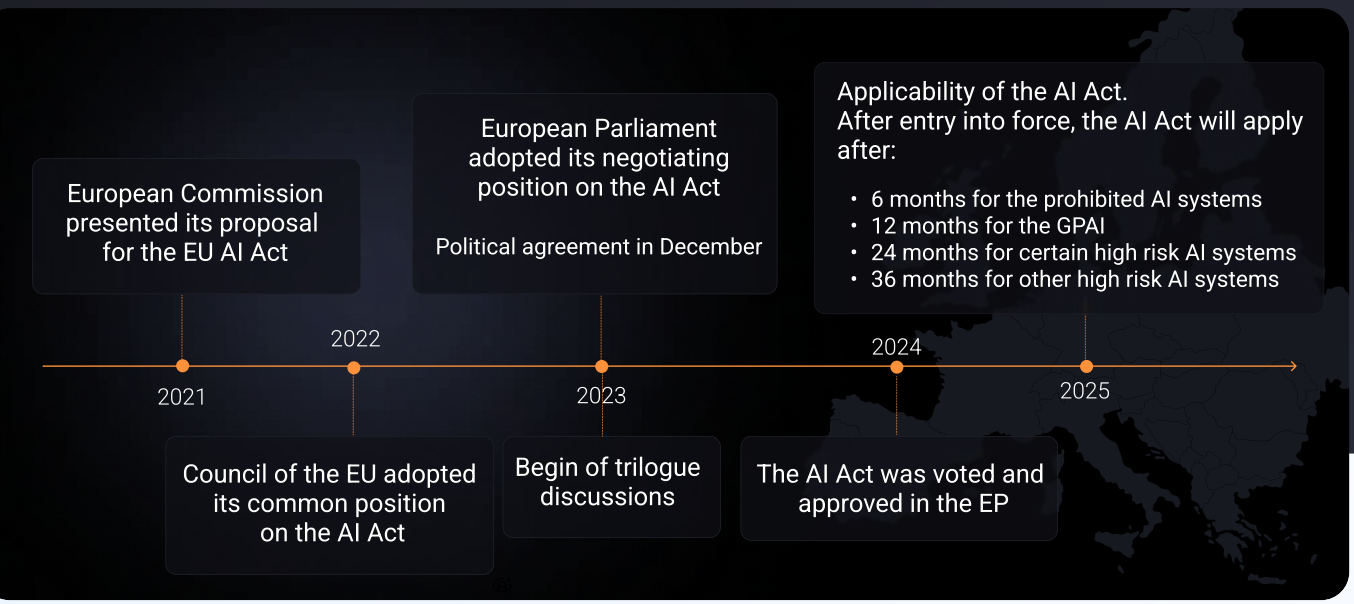
Timeline of the AI Act

In short, the AI Act is the first European-level regulation specifically aimed at overseeing AI. The journey to adopting the AI Act involved deliberations, negotiations, and revisions, resulting in a political agreement in December 2023. While a draft version of the AI Act is already available, the final text is expected to be published in the coming months. Once entered into force, there will be a transition period of 24 months before the provisions become fully applicable, with certain exceptions taking effect sooner at 6, 12, or later at 36 months.

Risk-based approach

At its core, the AI Act uses a risk-based approach, classifying AI systems into different risk levels based on their potential impacts on individuals and society. For instance, some AI systems pose too many risks and are consequently prohibited (think about manipulative or very intrusive systems). Other AI systems present high risks from an ethical perspective. For example, if the systems are not properly designed and developed, they could potentially lead to discrimination, violations of fundamental rights, or impact society or someone's right to privacy in a negative way. These AI systems are not prohibited per se, but measures should be implemented to identify and mitigate those risks. Then, there are AI systems with lower risks that still require specific measures to ensure transparency.

As indicated in the title, this white-paper will focus on high-risk AI systems. If you're unsure whether your (future) AI project falls into the high-risk category, go page 5 to discover the classification of high-risk AI systems.



Why think about the AI Act early?

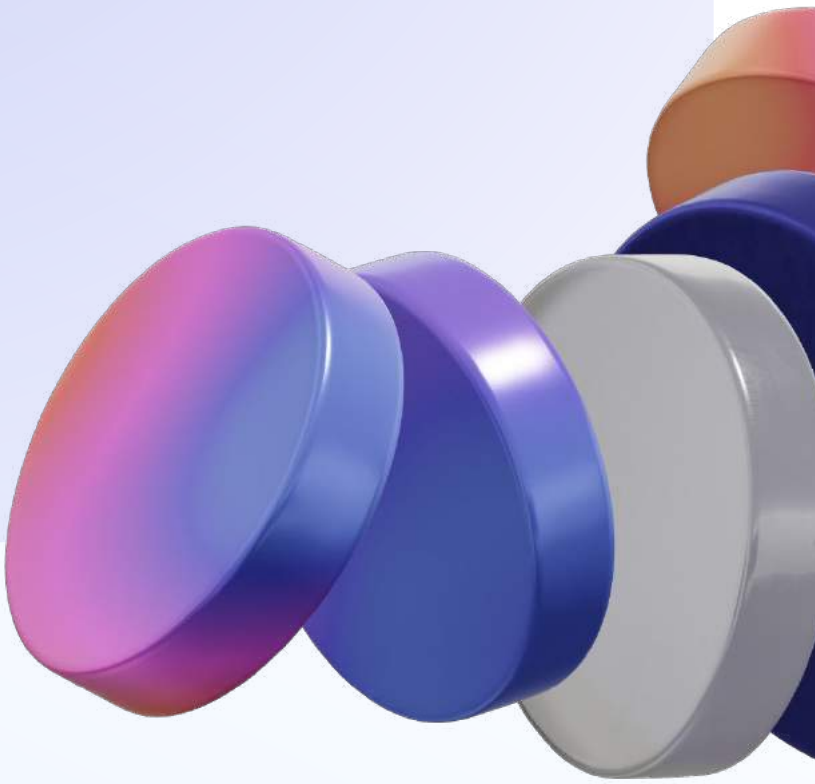
You might wonder why you should consider the AI Act now when most provisions won't be applicable for one or two years. At ML6, we see 3 main perspectives to consider:

From an ethical standpoint, adhering to the AI Act early on enables the development of responsible AI solutions. Viewing the AI Act just as an administrative burden overlooks its benefits. Proactively identifying and mitigating risks enhances the quality of your AI system. For example, this proactive approach helps identify and reduce biases to create a more fair AI solution prior to its development or its production release. The requirements also emphasise security best practices and transparency, fostering user trust in AI systems.

From a legal and financial standpoint, non-compliance with the AI Act, especially for high-risk AI systems, can lead to substantial fines - up to 35 million euros or 7% of your company's global revenue for major violations. Even minor infractions can result in fines of up to 7.5 million euros or 1.5% of turnover. Next to fines, non-compliance with the AI Act could lead to liability if the system would cause damages to individuals.

This situation reminds us of what happened with GDPR. Based on past lessons, we suggest getting ready for the AI Act before it becomes too urgent.

From a technical viewpoint, developing AI solutions is a time-intensive process. If you understand the AI Act requirements early on, you can integrate them into your project from the beginning. This makes compliance easier later and prevents the need to overhaul your technical setup or write extensive documentation when your AI system has already been developed two years ago. At ML6, if we identify that a client's project falls into the high-risk category, we inform and advise our client about the upcoming requirements. Our way of working also allows us to implement these requirements early on, such as conducting risk assessments, adopting a security-by-design approach, and closely collaborating with users throughout the development.



High-risk AI systems

It is well understood that AI systems can pose certain risks for individuals and the society. For example, an AI driven surgical robot could potentially hurt an individual, resulting in a negative impact on that person's health or safety. Next to the potential physical impact of AI systems, the use of an AI system could also lead to the violation of someone's fundamental rights, such as the right to privacy or the right to a fair trial. Imagine a company using AI to screen job applicants based on resumes, learning from past hiring data. If not trained properly on representative data, the AI system could unintentionally discriminate against candidates from underrepresented groups, violating their right to non-discrimination.

The AI Act tries to tackle these risks following a risk based approach. If a system poses an unacceptable risk, it will be prohibited. Other systems, posing high risks, are classified as "high-risk AI systems". These systems can still be developed and used, but they need to meet specific requirements to ensure ethical development and use, thereby minimising risks. You can find an overview of the requirements on page 10.

In this section, we will focus on the classification of high-risk AI systems identified by the AI Act. We will discuss (i) which systems are considered high-risk and (ii) the reason why these systems are classified as high-risk.

The high-risk AI systems are divided into two annexes (Annex II and Annex III of the AI Act), so that different rules can apply to the systems in the respective annexes. For example, only the systems listed in Annex III shall be registered in a database. Or only the deployers of a system listed in Annex III should perform a fundamental rights impact assessment. Another example: the rules for AI systems in Annex II will apply after 36 months after entry into force whereas for systems in Annex III the rules will apply after 24 months.

A. Annex II: Safety components of regulated products

The first category of high-risk AI systems are the so-called safety components of regulated products - or when the AI system is the regulated product as such. These products are already impacted by sector specific regulation. Examples include elevators, toys, high pressure devices, medical products, airplanes, trains, motor (vehicles), e-bikes, and the equipment and protective systems intended for use in potentially explosive atmospheres.

If the safety component of a product involves an AI system, it will be classified as high-risk due to its potential adverse impact on health, safety, or individuals' fundamental rights. Think about an autonomous robot for personal assistance, or a decision-support system in the healthcare sector. As the stakes for life and health are particularly high in these sectors, the AI systems should be able to operate safely, reliably, and with high accuracy.

If the AI system is not a safety component of regulated products, it will not be categorised as high-risk. For example, an AI system selecting elevator music would not be classified as high-risk since its failure would not pose risks to the health and safety of individuals.

B. Annex III: AI systems used in specific industries

1. Education

Some AI systems used in education are classified as high-risk because they can significantly influence someone's educational and professional course. Specifically, in the AI Act, four types of systems are identified as high-risk AI systems:

- AI systems used for determining access or admission to educational institutions, such as a system that decides whether you can be admitted to a specific school;
- AI systems used for evaluating learning outcomes of persons, such as a system that grades tests;
- AI systems used for assessing the appropriate level of education, such as a system that recommends suitable learning content to students;
- AI systems used for monitoring and detecting prohibited behaviour of students during tests.

These systems should meet the requirements for high-risk AI systems, as inadequate design and use could result in significant intrusiveness and potential violations of individuals' rights to education, training, as well as the right not to be discriminated.

2. Employment

Similar logic applies to AI systems in employment, given that these systems, like those used in education, can significantly influence individuals' future career trajectories.
During the recruitment process and in the evaluation, promotion, or retention of persons in work-related contexts, AI systems have the potential to perpetuate historical patterns of discrimination. Therefore, the following systems are seen as high-risk under the AI Act:

- AI systems used for the recruitment and selection of persons, such as a system used to screen job applicants based on resumes;
- AI systems used for making decisions affecting terms of the work related relationship, such as promotion and termination;
- AI systems used for allocating tasks based on individual behaviour, personal traits or characteristics.

Also, systems used for monitoring or evaluation of persons on the work floor are considered high-risk, as these may undermine the workers' fundamental rights to data protection and privacy.

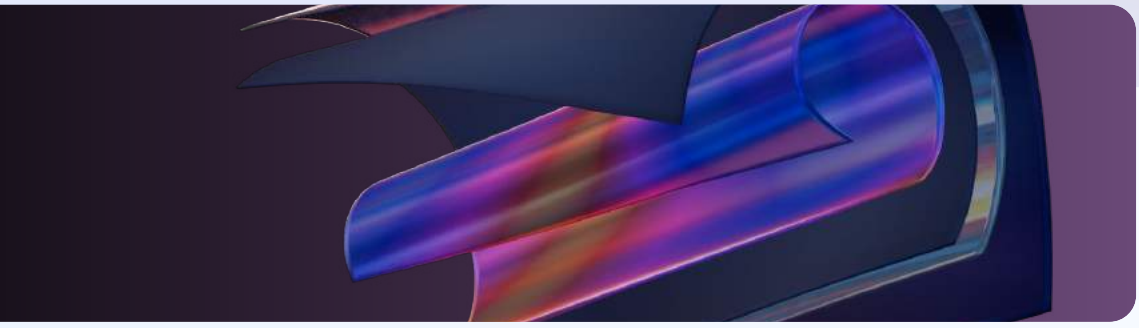
3. AI systems intended for the administration of justice

AI systems used to research, interpret facts and the law, and to apply the law to a concrete set of facts are considered high-risk. For example, an AI system that assesses arguments in legal proceedings and generates judgments could significantly impact the concerned individuals' right to an effective remedy and to a fair trial.

Additionally, this category includes AI systems used to influence election or referendum outcomes or the voting behaviour of individuals, such as political chatbots. These systems are categorised as high-risk due to their potential impact on democracy and the rule of law.

However, AI systems used solely for ancillary administrative tasks that do not affect the actual administration of justice in individual cases, such as anonymisation or pseudonymisation of judicial decisions are not classified as high-risk AI systems.

AI systems can significantly influence individuals' future career trajectories.



4. Access to and enjoyment of essential public services and benefits

Another category of high-risk systems includes those that impact people's access to and enjoyment of essential public services and benefits.

This category includes:

- AI systems used for determining whether certain benefits and services should be granted, denied, reduced, revoked or reclaimed by authorities (such as social services providing protection in cases of maternity, illness, industrial accidents, and loss of employment);
- AI systems used for evaluating someone's creditworthiness, for example when applying for a loan;
- AI systems used for risk assessment and pricing in relation to natural persons for health and life insurance;
- AI systems used for evaluating and classifying emergency calls by natural persons or establishing priority in the dispatching of emergency first response services, including by police, firefighters, or in case of medical emergency. Such systems are high-risk since they make decisions in very critical situations for the life and health of persons and their property.

If such systems were not properly designed, developed and used, these could infringe the involved peoples' fundamental rights and could lead to serious consequences, including financial exclusion and discrimination.

5. AI systems in critical infrastructure

Safety components used in the management and operation of critical digital infrastructure - such as AI systems supporting road traffic and essential utilities, like water, gas, heating and electricity - are inherently high-risk.

Likewise, AI systems designed to directly protect the physical integrity of critical infrastructure or health and safety of persons and property, such as water pressure monitoring or fire alarm controls in cloud computing centres, are high-risk AI systems.

This classification is due to the potential consequences of failure or malfunctioning of these systems, which could endanger the lives and health of many individuals and result in significant disruptions to social and economic activities.

6. Migration, asylum and border control management

In the context of migration, there are also high-risk AI systems, namely:

- AI systems used for assessing certain risks posed by natural persons entering the territory of a Member State or applying for visa or asylum.
- AI systems used for assisting authorities for the examination, including related assessment of the reliability of evidence, of applications for asylum, visa and residence permits and associated complaints.

These AI systems affect people who are often in a particularly vulnerable position and rely on the outcome of the actions of the competent public authorities.

7. Law enforcement

Some AI systems used in law enforcement are classified as prohibited AI systems, including those used for real-time remote biometric identification in publicly accessible spaces for the purpose of law enforcement. It's important to note that there are exceptions, such as in cases involving threats to the life or physical safety of natural persons or terrorist attacks.

Other AI systems used for law enforcement purposes are not prohibited but are classified as high-risk systems. Think about an AI system checking through cameras whether or not someone is calling while driving.

The systems include:

- AI systems used for assessing the risk of a natural person to become a victim of criminal offences.
- AI systems used for the evaluation of the reliability of evidence in the course of investigation or prosecution of criminal offences.

Where such systems are not sufficiently transparent, explainable and documented, the exercise of individuals' important procedural fundamental rights - including the right to an effective remedy and to a fair trial as well as the right of defence and the presumption of innocence - could be hampered.

The AI system may also single out people in a discriminatory, incorrect or unjust manner, especially if the system is trained with low-quality data, and fails to meet adequate standards in terms of performance, accuracy or robustness.

8. Remote biometric identification systems

A final category of high-risk systems are remote biometric identification systems, which are defined as AI systems for the purpose of identifying natural persons without their active involvement, typically at a distance through the comparison of a person's biometric data with the biometric data contained in a reference database.

Technical inaccuracies can result in biassed outcomes and discriminatory effects, which is especially pertinent in relation to age, ethnicity, race, sex, or disabilities.

However, systems designed solely for authentication and identity confirmation, such as the face ID application on your phone, are not considered high-risk AI systems.

How to deal with the requirements for high-risk AI systems?

So what should you do if your AI project falls into the high-risk category?

In this section, we provide our general impressions of the requirements, go deeper into the specifics of these requirements, and share our experience on proactively implementing them.

General impressions

Chapter 2 of the current AI Act version describes the seven requirements for high-risk AI systems. Our general impression is that, while the legal aspects of the AI Act have been well-considered, the precise technical implementation of these requirements remains somewhat unclear. Terms like "judged to be acceptable", "as far as technically feasible" and "where appropriate" suggest that further clarification and practical guidelines will be necessary to minimise subjective interpretation. Another observation is that most requirements focus on enhancing transparency in AI systems, such as documenting the development process, providing user instructions, and implementing logging mechanisms to record AI system outputs. Transparency plays an important role in building trust in AI among individuals and society. Additionally, the extensive documentation and logging requirements also serve the purpose of enabling authorities to verify compliance with the regulation.

Requirements in a nutshell

1. Risk Management System

To fulfil the first requirement on risk management system, you need to establish a continuous and iterative process for identifying and mitigating risks associated with your project. This involves identifying foreseeable risks under intended use and potential misuse scenarios, implementing risk management measures, and testing the effectiveness of these measures. The continuous and iterative nature of this requirement is very important. You can start with an initial risk assessment analysis at the start of your AI project using frameworks such as the Assessment List for Trustworthy AI. However, it's important to re-evaluate those assessments if there are any changes in the project scope, such as adding new data sources or involving users with diverse backgrounds or interests. Throughout the development lifecycle, it's also important to pay attention to new insights that may impact the risks analysis, as discoveries during the development process may reveal additional non-identified risks that you need to address.

2. Data governance

The second requirement of the AI Act focuses on data governance, emphasising the importance of high-quality data when developing AI applications. But what exactly does "high-quality data" mean? It means ensuring that your training, test, and validation datasets are relevant, representative, complete, and as error-free and unbiased as possible. To achieve this, there are several data governance best practices to follow. For instance, to mitigate biases in your dataset, start by understanding the business context and analyse your data for biases. Techniques like creating a balanced dataset can help reduce bias. In the context of the AI Act, it's important to document your data governance practices, including data preparation, validation, and monitoring steps.

3. Technical documentation

As mentioned in the previous paragraph, documentation plays an important role in the light of the AI Act. Documentation not only promotes transparency but ensures compliance with regulatory requirements. In Annex 4 of the AI Act, you will find a detailed list of information that must be included in your technical documentation, covering aspects such as the AI system's goals, architecture, capabilities, limitations, and processes. If you're dealing with multiple high-risk AI systems or plan to do so, we strongly recommend creating templates with predefined sections outlining the required information. It's important to note that while AI developers often write the technical documentation, interpreting and implementing AI Act requirements may benefit from legal expertise. Consider seeking external advice to effectively translate regulatory requirements into practice.

4. Record-keeping

With record-keeping, or what we also call logging, you need to automatically record events of your AI systems to ensure traceability. Those logs can be used to identify potential risks and to monitor the system once it has been released into production. By leveraging detailed logs, you can effectively track system behaviour, diagnose issues, and maintain accountability throughout the lifecycle of your AI system.

5. Human oversight

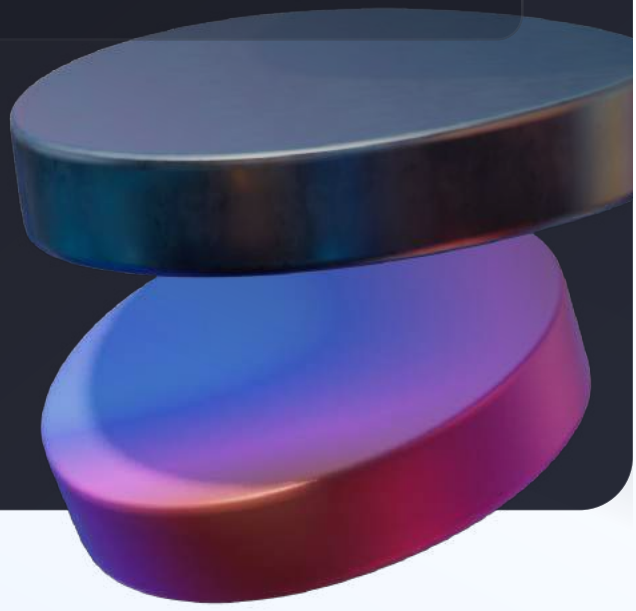
Human oversight involves ensuring that natural persons have control and oversight when needed. This means having a responsible individual who ensures that the system functions correctly as intended and can intervene to prevent unintended actions. It's important to assign this role to someone with the appropriate responsibilities and training to understand the AI system's limitations and know when intervention is necessary.

6. Transparency

Transparency in AI involves ensuring that users understand the capabilities and limitations of the system. This means providing clear details and instructions on how to use the AI effectively. By doing so, users can make informed decisions and use AI responsibly. Our recommendation is to directly engage with users during the development process of your AI system. Consider organising training sessions to explain how the system works, potential risks like overreliance or bias, and how to handle any issues that come up during use. Offering a basic introduction to AI can also help users feel more comfortable and confident using the technology responsibly.

7. Accuracy and cybersecurity

Accuracy and cybersecurity are important aspects concerning the performance and security of AI systems. Using metrics to measure performance ensures that the AI system works as intended, although accuracy alone shouldn't be the only metric used. It's important to consider a range of metrics to get a full picture of performance (think about fairness metrics). In addition to performance, you need to protect your system against different types of attacks. This includes safeguarding against common software attacks and specific vulnerabilities such as prompt injections that can affect AI systems.



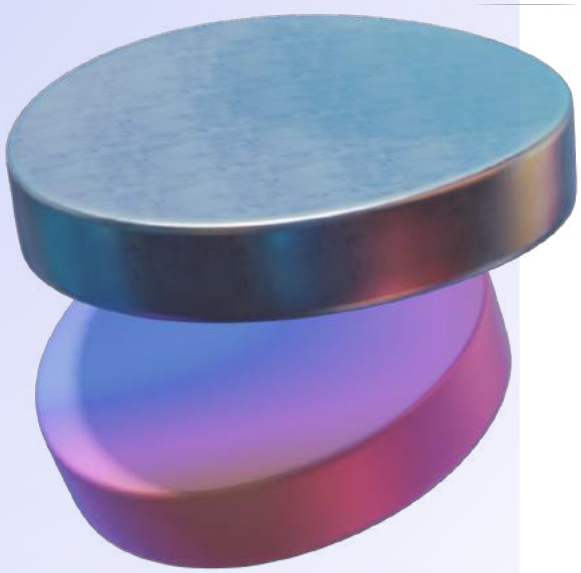
Disclaimers

Disclaimer 1

This white-paper is not based on the officially published text of the AI Act but on the texts circulating at the time of writing the blogpost (May 2024) - the final text may still (albeit slightly) differ from the used versions.

Disclaimer 2

The law provides for some exceptions regarding high-risk systems in Annex III (Article 6.2a of the AI Act) - for example, the system is not high-risk (even though included in Annex III) if the system does not pose a significant risk of harm to the health, safety, or fundamental rights of natural persons, including by not materially influencing the outcome of decision-making. Since we still want to await the concrete implementation of that exception, and because we prefer to err on the side of caution, especially in certain grey areas, we have not delved deeper into these exceptions in this white-paper.



Disclaimer 3

The white-paper is or contains no legal advice - to the best of our ability, we have based ourselves on the wording of the current version of the AI Act (the articles and the recitals) - however, some passages or examples are merely our interpretation and may differ from any subsequent guidelines, case law, or opinions.

ML6 , building AI solutions you can trust

ML6 leads in responsible AI innovation, embedding ethics in solutions from advising on the EU AI Act to creating transparent, fair, and accountable AI. Our mission is to empower businesses with advanced AI that meets the highest ethical and legal standards, ensuring AI you can trust.

We systematically identify, mitigate and monitor potential ethical and legal risks in your projects. We proactively advise you on the upcoming legal requirements of the AI Act. This way, we aim to maximise benefits while proactively minimizing potential risks.

Feel free to schedule a meeting with our expert team.

Pauline Nissen
Ethical AI Lead

Contact »

Michiel Van Lerbeirghe
Legal Counsel

Contact »