Navigating European AI Regulation: The EU AI Act and Principles for Trustworthy AI

KILAB

trail

Content

Ladies and Gentlemen,

In recent times, we have witnessed the potential of artificial intelligence in all industries. We are deeply engaged with these dynamic developments, and use our knowledge and expertise to support companies in the efficient adoption and implementation of AI in their business practices.

As AI continues to evolve,
the importance of developing ethical and trustworthy systems has become increasingly clear. This is not just a matter of best practice but a legal imperative, as emphasized by the EU AI Act.
This landmark regulation establishes a robust legal framework, mandating that AI systems adhere to strict standards of safety, transparency, and accountability.

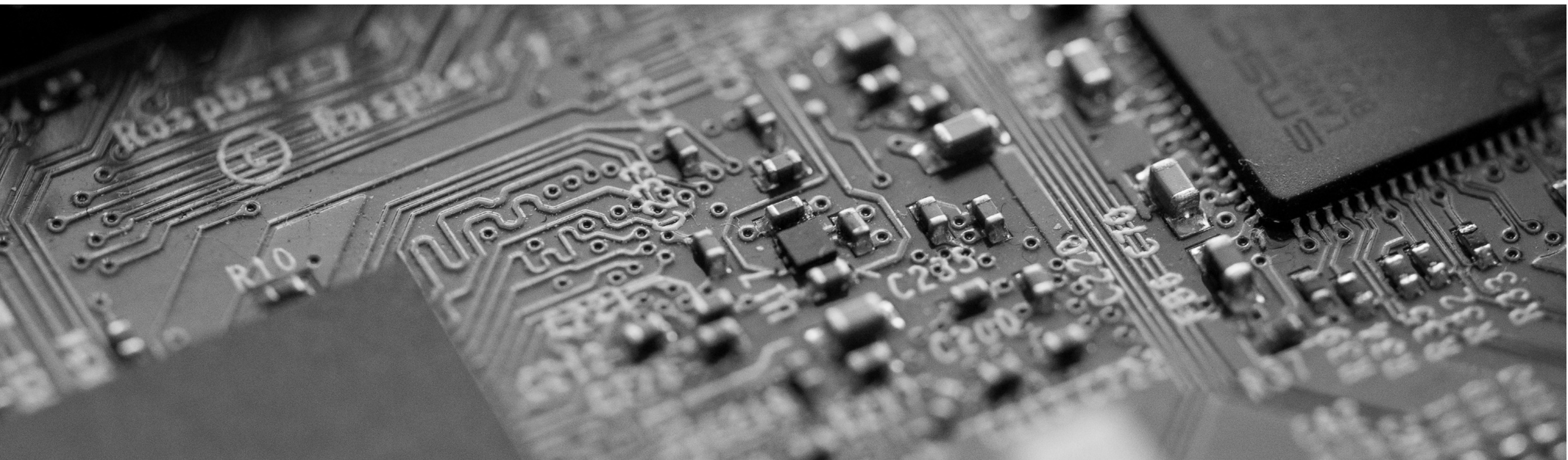
At our KI-Lab, we are dedicated to helping companies navigate this complex regulatory landscape. With our close connection to the Technical University of Munich, we stay at the forefront of AI research and combine it with practical strategies to accelerate AI integration across various sectors. Through an in-depth exploration of the EU AI Act and its implications, we aim to provide guidance on implementing AI that is not only innovative but also ethical and compliant with the highest standards of trustworthiness.

We look forward to an exciting journey into the future of trustworthy AI!

Best regards,

Univ.-Prof. Dr. Dr. h. c. mult.
Horst Wildemann

Horst Wildemann studied mechanical engineering and business administration in Aachen and Cologne, obtained his doctorate in 1974 and habilitated at the university of Cologne in 1980. He has been a professor of business administration at the Technical University of Munich since 1989 and has been appointed to numerous prestigious universities. He was inducted into the Logistics Hall of Fame in 2004 and has remained active in research and teaching as well as in his role as Managing Director at TCW since his emeritus status in 2010. His scientific work includes over 40 books and 600 articles, with a focus on corporate management, production and logistics.



Executive Summary

The rapid evolution of Artificial Intelligence (AI) has brought about both groundbreaking advancements and serious risks. Incidents, such as misinformation spread by chatbots and biases inherent in certain AI models, have underscored the urgent need for ethical AI practices and stringent regulations.

The EU AI Act, effective from August 2024, represents a pivotal step towards addressing these challenges by establishing comprehensive, legal guidelines to ensure the ethical and trustworthy deployment of AI systems.

Key Points

- **Urgency of AI Regulation:**
  The complexity and opacity of AI decision-making demand immediate regulatory action to prevent errors and adversarial attacks. The EU AI Act establishes a legal framework to ensure AI systems are lawful, robust, and ethical, protecting human rights and safety.

- **Ethical Principles and Practical Implementation:**
  While various principles for ethical AI have been published, operationalizing them in business and research has been challenging. The EU AI Act aims to bridge this gap by enforcing practical measures.

- **The EU AI Act:**
  The world's first comprehensive AI law, the EU AI Act categorizes AI systems by risk levels (unacceptable, high, limited, and minimal risk) and imposes specific requirements for high-risk systems. These include robust data governance, continuous risk management, technical documentation, human oversight, and cybersecurity measures. The Act also introduces regulations for General Purpose AI (GPAI) models, which are adaptable AI models capable of performing a wide range of tasks. Due to their flexible nature, GPAI models, such as ChatGPT, pose unique regulatory challenges. The EU AI Act mandates that GPAI providers maintain up-to-date technical documentation, ensure compliance with Union law on copyright, and provide necessary information to downstream providers to ensure transparency and accountability.

- **Challenges and Criticism:**
  The AI Act has faced criticism for potentially stifling innovation and competitiveness in Europe, as well as for the complexity and ambiguity in risk classification which lead to uncertainties that could deter AI adoption and investment. To address these concerns, the EU AI Act includes measures in support of innovation, such as AI regulatory sandboxes. These controlled environments allow startups and SMEs to develop, test, and refine AI systems before market launch, facilitating innovation while ensuring compliance.

- **Call to Action:**
  Organizations must prepare for the AI Act by ensuring compliance with its requirements. Strategic consulting and technological solutions, such as those offered by the KI-Lab and trail, can help companies navigate the new regulatory landscape effectively.

In conclusion, the EU AI Act is a crucial step toward responsible AI innovation, balancing the potential benefits of AI with the necessity of ethical and safe deployment. By addressing current uncertainties and fostering a culture of compliance, the Act aims to pave the way for a future where AI advancements are achieved responsibly and transparently.

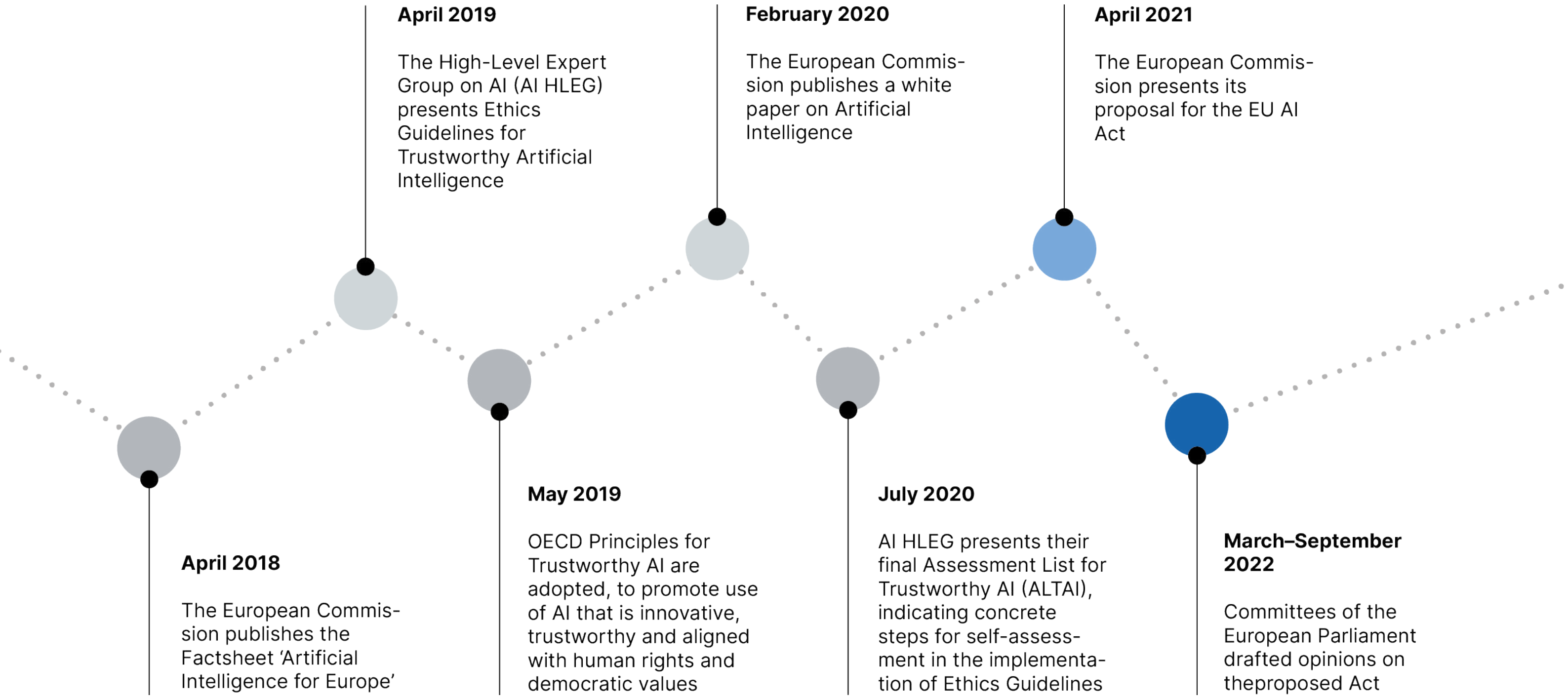
Introduction

Artificial Intelligence (AI) has proven to be a double-edged sword, bringing significant advances but also posing serious risks. Since the release of ChatGPT in November 2022, the media has highlighted the potential risks associated with the inappropriate use of AI, with a focus on Generative AI (GenAI). Incidents have surged at an alarming rate, ranging from misinformation spread by chatbots, such as Air Canada's chatbot falsely promising a discount to a passenger[1], to racial biases in language models that unfairly graded Asian American students lower[2], or Google's AI spreading misleading or at times dangerous information such as adding glue to pizza[3].

These alarming cases underscore the potential harmfulness of AI and foster a widespread sense of urgency and a call for ethical AI and stringent regulations – a call that was answered by both the development of ethical principles and guidelines, as well as by integrating these into legislation, namely through the EU AI Act.

Regulating emerging technologies is not new; it dates back to the steam engine. However, AI is widely perceived as one of the most disruptive technologies of our time, amplifying the urgency for control and regulation.[4] Despite some expectations that this might impede AI's future development, the prevailing sentiment

Figure 1: Timeline of the developments in Trustworthy AI including principles and guidelines (gray) as well as the EU AI Act (blue).



April 2018

The European Commission publishes the Factsheet 'Artificial Intelligence for Europe'

April 2019

The High-Level Expert Group on AI (AI HLEG) presents Ethics Guidelines for Trustworthy Artificial Intelligence

May 2019

OECD Principles for Trustworthy AI are adopted, to promote use of AI that is innovative, trustworthy and aligned with human rights and democratic values

February 2020

The European Commission publishes a white paper on Artificial Intelligence

July 2020

AI HLEG presents their final Assessment List for Trustworthy AI (ALTAI), indicating concrete steps for self-assessment in the implementation of Ethics Guidelines

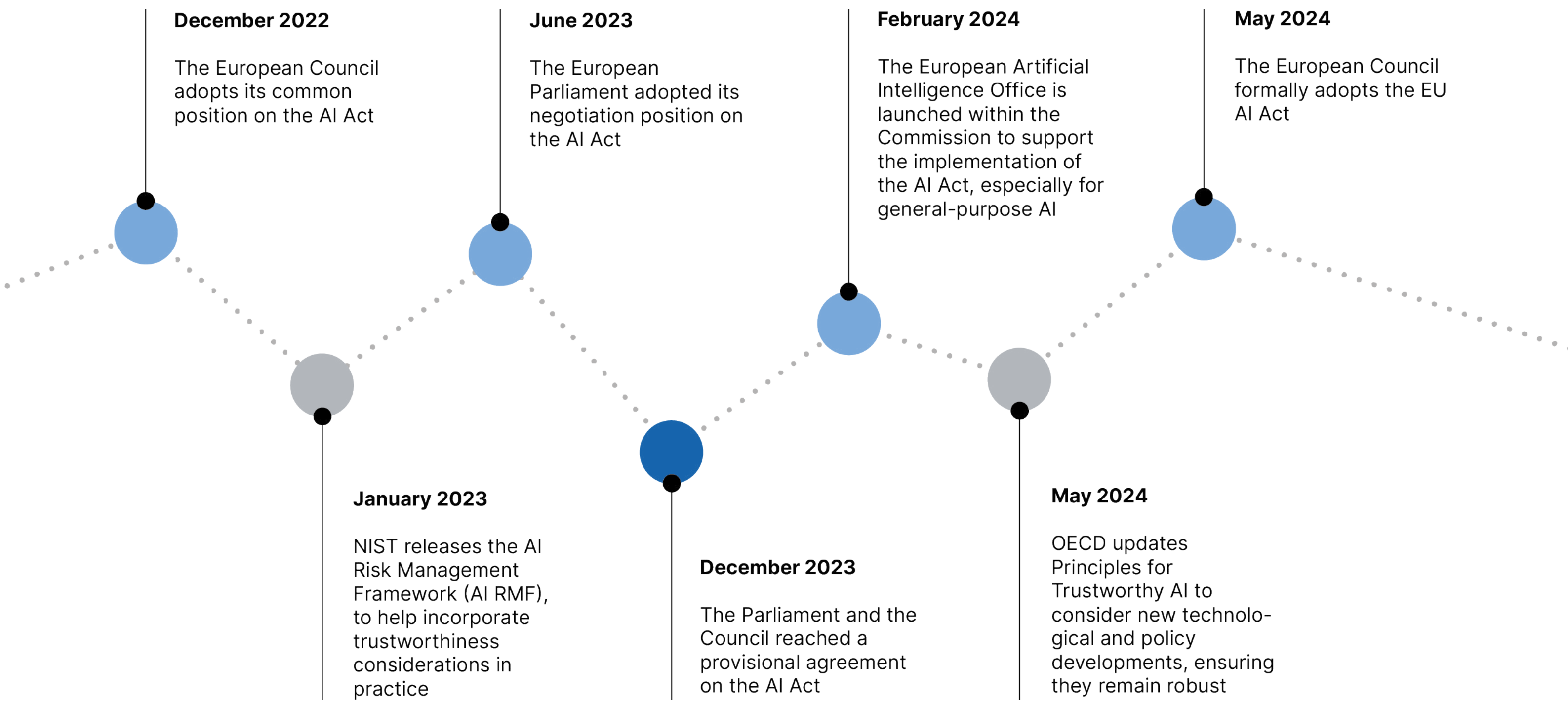
April 2021

The European Commission presents its proposal for the EU AI Act

March–September 2022

Committees of the European Parliament drafted opinions on theproposed Act emphasizes the necessity of regulation to ensure ethical use, minimize bias, foster transparency, protect human rights and safety, and establish accountability as AI continues to evolve. Before 2021, the AI landscape was primarily ruled by abstract guidelines and principles.

Buzzwords like ethical AI and trustworthy AI, were conceptualized through sets of ethical principles or checklists, but practical implementation and operationalization were largely lacking. This gap between theory and practice remains one of the key challenges to date, compounded by scarcity of concrete knowledge about AI trustworthiness and slow or nonexistent established processes for implementation. However, the introduction of the EU AI Act, along with other regulations in the EU and the US, and the growing availability of both free resources and visionary start-ups–specifically focused on ensuring compliance with the EU AI Act point towards a promising direction for future AI governance. With the latest release and final adoption of the EU AI Act, concerns about how to lawfully implement its rules are increasingly being raised in forums and conferences by top executives as well as AI developers.



December 2022

The European Council adopts its common position on the AI Act

January 2023

NIST releases the AI Risk Management Framework (AI RMF), to help incorporate trustworthiness considerations in practice

June 2023

The European Parliament adopted its negotiation position on the AI Act

December 2023

The Parliament and the Council reached a provisional agreement on the AI Act

February 2024

The European Artificial Intelligence Office is launched within the Commission to support the implementation of the AI Act, especially for general-purpose AI

May 2024

OECD updates Principles for Trustworthy AI to consider new technological and policy developments, ensuring they remain robust

May 2024

The European Council formally adopts the EU AI Act

But why is the need for AI regulation so urgent? The decision-making process of AI systems and, most importantly, the development of decision rules is often intransparent. This implies that the decisions made by algorithms cannot always be retraced and understood, and sometimes don't adhere to logical reasoning, leading to unexpected errors in seemingly simple tasks and vulnerability to adversarial attacks[5][6]. This lack of transparency is partly due to the increasing complexity of AI systems. While this complexity is not inherently negative, it poses a challenge to understanding and controlling these systems as they continue to evolve with the rapid pace of technological innovation. The goal is to make AI systems trustworthy, leading to lawful, robust, and ethical AI[7]. Establishing both normative and legal frameworks is crucial for the safe and ethical development of AI systems, ensuring that humans remain in control.

Ethical AI and the Jungle of Principles

One response to increasing risks related to AI systems was the call for "ethical AI". This call led to the publication of AI principle guidelines by major institutions, organizations, and governments. These principles are abstract, high-level ethical rules that AI systems, as well as their developers and deployers, should adhere to.

At time of writing, more than 200 different sets of principles have been published[8]. Some of the most prominent and influential sources include the OECD's Recommendation of the Council of Artificial Intelligence[9], IEEE's Ethically Aligned Design (Version 2)[10], and the European Commission's Ethics Guidelines for Trustworthy AI (AI HLEG)[11].

With the proliferation of publications on AI ethics, reviews have emerged that provide a comprehensive overview of the differences in definition, application, and reasoning across these principles[4][8][12].

Given the large overlap between the different sets of principles, the most prominent and frequently cited ones can be summarized as follows:

1. Transparency and explainability:
Developers should increase transparency in the underlying data during design and development, and implement explainable algorithms to ensure the decision-making processes of their AI systems are understandable.

2. Security and safety:
AI systems should not be accessible to unauthorized entities to prevent misuse and minimize harm.

3. Robustness and reliability:
AI systems should produce accurate results under different conditions.

4. Justice, fairness and non-discrimination:
AI systems should not discriminate against any groups ensuring fair decision-making processes.

5. Privacy:
individuals should have full control over their personal data, and AI systems must respect user privacy.

6. Accountability and responsibility:
AI systems should have clearly defined stakeholders who are accountable for their actions and decisions.

Additionally, two more overarching ethical principles are often included, namely **beneficence (also non-maleficence)** which states that AI systems should promote well-being and be motivated by **human good**, and **human autonomy** and **human oversight** (also **human-at-center approach**) which aims to place humans in full control and equip them with a full understanding of the system and its decision-making processes.

However, one critique commonly mentioned in both past and current literature is the lack of practical applications of such guidelines. Abstract and non-tangible ethics are well-defined, and their importance is undoubted, yet their implementation in business and research settings is often left for the developers and deployers to decide. This leaves room for interpretation, decreases standardization of AI practices,and creates opportunities for 'whitewashing,' superficial ethical compliance, and lack of transparency.

Moreover, some of these principles are not AI-specific but rather general principles that could apply to various technological advances, such as water boilers[4]. This lack of precision hinders concrete operationalization, as the implications of such principles might vary between different technologies.

Practical Examples

Despite these challenges, the growing body of available online resources is a good sign we're heading in the right direction. Various organizations offer educational content about AI principles and trustworthy AI, as well as checklists and a catalog of tools and metrics for trustworthy AI, aiming to bridge the gap between research and practice.

The OECD provides an overview of the principles and recommendations, which are based on the Recommendation of the Council on Artificial Intelligence (OECD/LEGAL/0449), and provides further information on trustworthy AI, making it an accessible and comprehensive resource. Note, however, that the OECD focuses on guiding policy makers rather than AI developers and deployers.

Additionally, the High-Level Expert Group on Artificial Intelligence (AI HLEG) published several deliverables, including the Policy and Investment Recommendations for Trustworthy AI, a final Assessment List for Trustworthy AI (ALTAI) and Sectoral Considerations on the Policy and Investment Recommendations.[13] The ALTAI, in particular, attempts to bridge the gap between research and application by providing a web-based tool, alongside a traditional report, allowing users, developers and deployers of AI systems to work with the checklist interactively. This checklist aims to operationalize their proposed seven key requirements and offer practical implementation guidelines along with best practices[14]. While the implementation recommendations are still very abstract and the operationalization is lacking, ALTAI is a good first step, albeit with room for improvement in terms of its practical recommendations.

In the United States, institutions like NIST and the newly established Institute for Trustworthy AI in Law and Society (TRAILS)[15], which recently received a $20 million grant, also emphasize the importance of trustworthy and ethical AI principles. They are 'focused on transforming the practice of AI from one driven primarily by technological innovation to one that is driven by ethics, human rights, and input and feedback from communities whose voices have previously been marginalized'[15]. Specifically designed for risk mitigation, NIST published the AI Risk Management Framework (AI RMF)[16] and an accompanying playbook. The playbook aims to provide suggestions of actions

to achieve the outcome laid out in the AI RMF Core. Although AI RMF defines how to govern, map, measure, and manage AI Systems, this framework, too, lacks practical implementation guidance.
In conclusion, it is evident that various theoretical formulations are in place, such as ethics guidelines and formal regulations, but the gap between theory and practice has still not been bridged, and ethical principles are often not in line with regulatory requirements.

Norms and Standards as a Middle Ground Between Guidelines and Principles

One strategy to find a middle ground between loose principles and strict regulation is the implementation, distribution and adaptation of official norms or standards, such as ISO/IEC norms. Such standards not only offer general advice to ensure trustworthiness of AI systems but give more practical steps for application-specific contexts. Consequently, they enable clear auditing structures and facilitate standardization. An audit or certification based on these well-known norms can boost business sales, ensure compliance and communicate responsible AI efforts.

However, AI-specific norms and standards are still in development, and the integration of EU AI Act requirements is mostly lacking. The **ISO/IEC 42001:2023**, for instance, is a great start for establishing an AI management system, but is not fully in accordance with the EU AI Act, covering only a few aspects required under the new regulation. Keeping specific support regarding norms and standards up to date is challenging, due to the ever-evolving landscape. Consulting companies and organizations and tools focusing on AI-governance are valuable additions to a trustworthy AI strategy.

Additionally, the business model of certifying AI systems is gaining popularity. Organizations such as TÜV, DEKRA, and LNE are expanding their business models to provide companies deploying AI with official audits. This means AI systems can undergo audits by external companies, similar to existing IT security practices, enhancing user safety and transparency. However, several issues, such as liability, costs, and administrative overhead, remain unresolved.

One Step Towards Broad AI Responsibility: The EU AI Act

One step to bridge the gap from principle to practice is embodied in building a legal framework and embedding AI in a regulated, legislated space, both nationally and internationally. The European Commission was the first one to start such an endeavor by proposing the EU AI Act[17], which regulates AI systems – or rather their application – according to their risk levels. First presented in April 2021, the EU AI Act[18] has been revised multiple times to stay up to date with the ever-advancing technological innovations and development, such as the widespread implementation and use of large language models, image generation and more, with the most prominent example being ChatGPT, which is now ubiquitous. Used correctly, such breakthroughs can be of great benefit and open a world of opportunities. However, like any advanced technology, they can also be deployed irresponsibly, unethically or malevolently, requiring effective regulation. The EU is trying to balance innovation and the risks that arise through this new technology by fostering the responsible use of AI systems with their digital strategy and the AI Act.

In what follows, let us have a closer look at the AI Act and what exactly it currently entails. The final text of the EU AI Act was recently adopted and has come into force as of August 2024, leading to a two-year grace and transition period before it comes into full effect. Specifically, after

- **6 months** latest, organizations have to comply with the regulation on prohibited AI systems and AI literacy
- **12 months** latest, organizations have to comply with the regulation on general purpose AI systems (GPAI)
- **24 months** latest, organizations have to comply with all other regulations (i.e. on high-risk or limited risk AI systems).

There is an extended period for providers or deployers of high-risk AI systems in certain applications that are already subject to other legislation in the European Union (especially those which are regulated under the "New Legislative Framework"). This includes AI systems that are used as a safety component in already regulated products, such as machinery, lifts, protective equipment or toys (see Annex I of the AI Act). Organizations offering solutions in these application areas need to comply with the EU AI Act latest 36 months after enactment.

A more extensive timeline of the AI Act is shown in Figure 2.

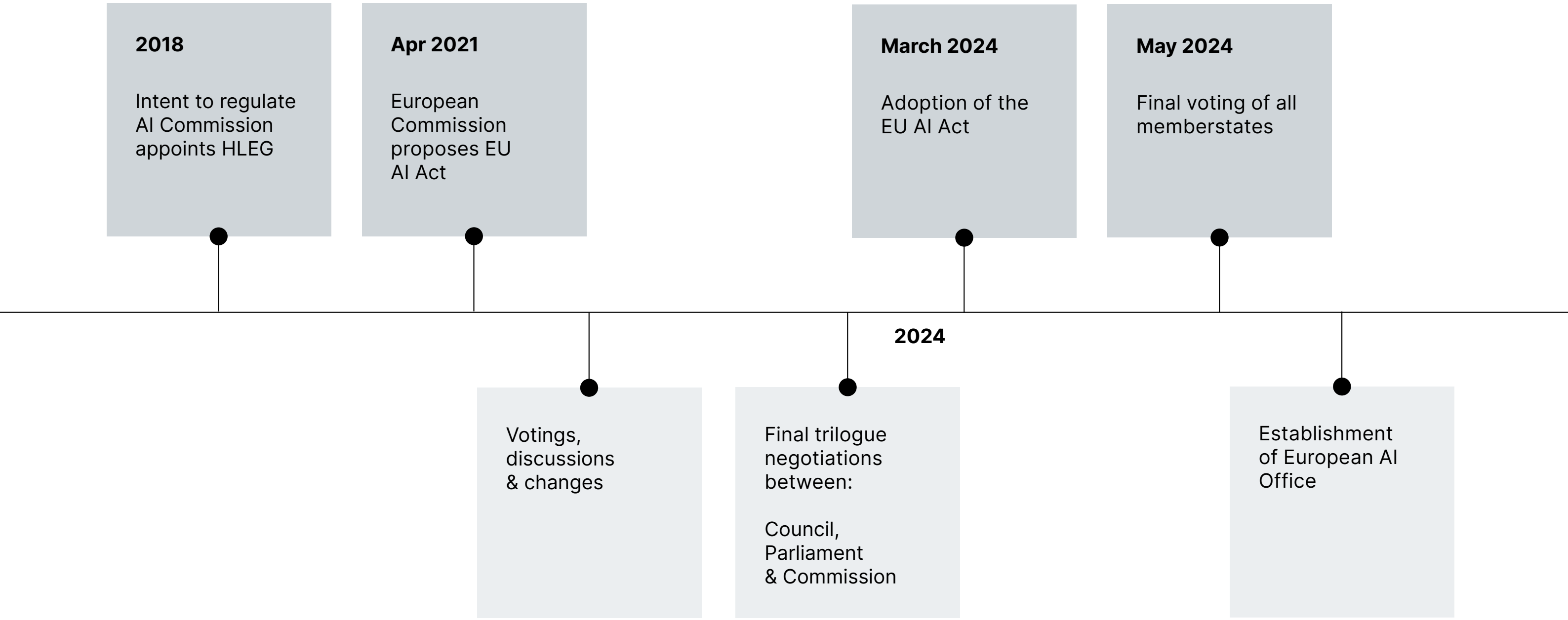

2018
Intent to regulate AI Commission appoints HLEG
Apr 2021
European Commission proposes EU AI Act
Votings, discussions & changes
Final trilogue negotiations between:
Council, Parliament & Commission
2024
March 2024
Adoption of the EU AI Act
May 2024
Final voting of all memberstates
Establishment of European AI Office

Figure 2: Timeline of the EU AI Act.

A major point of debate during the negotiations of the AI Act has been the definition of an AI system. In article 3, paragraph 1, the AI Act now defines an AI system as 'a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments'.

Other important definitions as per the article 3 of the AI Act include:

- **Provider:** 'a natural or legal person, public authority, agency or other body that develops an AI system or a general-purpose AI model or that has an AI system or a general-purpose AI model developed and places it on the market or puts the AI system into service under its own name or trademark, whether for payment or free of charge' (paragraph 3)

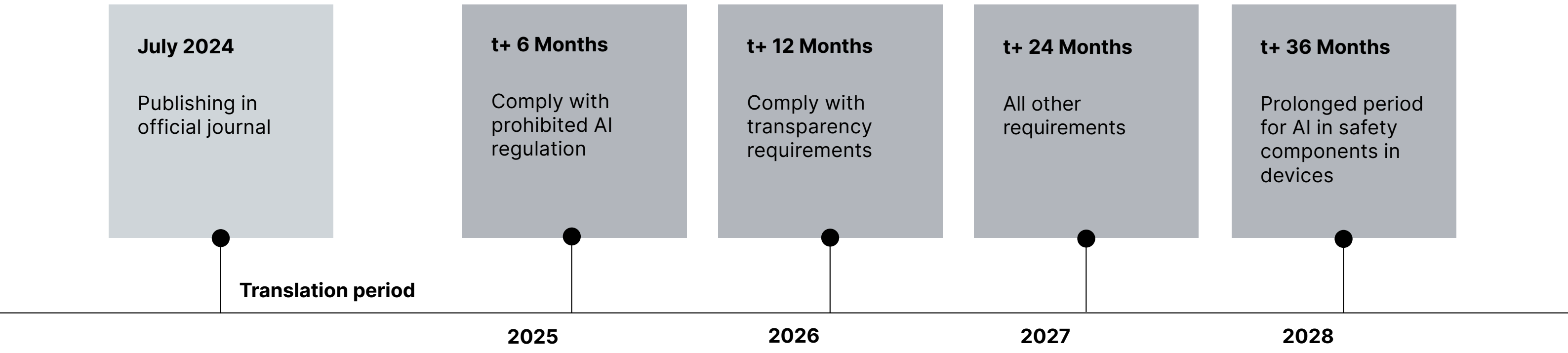
- **Deployer:** 'a natural or legal person, public authority, agency or other body using an AI system under its authority except where the AI system is used in the course of a personal non-professional activity' (paragraph 4)

Consider a startup that develops an innovative AI system for credit scoring. Banks buy this software and implement it to manage loan distribution. In this scenario, the startup is the provider, and the banks are the deployers. The AI Act defines different obligations for providers and deployers which we will discuss further below.

Further terms include:

- **Importer:** 'a natural or legal person located or established in the Union that places on the market an AI system that bears the name or trademark of a natural or legal person established in a third country' (paragraph 6)

- **Distributor:** 'a natural or legal person in the supply chain, other than the provider or the importer, that makes an AI system available on the Union market' (paragraph 7)



July 2024

Publishing in official journal

Translation period

t+ 6 Months

Comply with prohibited AI regulation

2025

t+ 12 Months

Comply with transparency requirements

2026

t+ 24 Months

All other requirements

2027

t+ 36 Months

Prolonged period for AI in safety components in devices

2028

An importer or distributor can be considered an AI's provider if it puts its own trademark on a product or substantially changes the product. In this case, the importer or distributor must assume all responsibilities for providers as outlined in the Act.

As mentioned, the AI Act classifies AI systems by their risk, i.e. according to the likelihood and severity of the damage they could inflict to health, safety, or fundamental human rights. There are four risk categories: unacceptable, high, limited, and minimal risk, which are depicted in Figure 3. The higher the risk of a system, the stricter are its regulatory requirements. Notably, AI systems which are used exclusively for military or defense purposes are exempt from these requirements, as are AI systems solely used for research and innovation. The AI Act does also not apply to individuals using AI for non-professional purposes.

AI Act related definitions

Provider

Develops or commissions an AI system and markets or deploys it under their name or trademark.

Deployer

Uses an AI system under their authority, excluding personal, non-professional use.

Importer

A person or entity in the Union that markets an AI system from a third country.

Distributer

A supply chain entity that makes an AI system available in the Union market, not being the provider or importer.

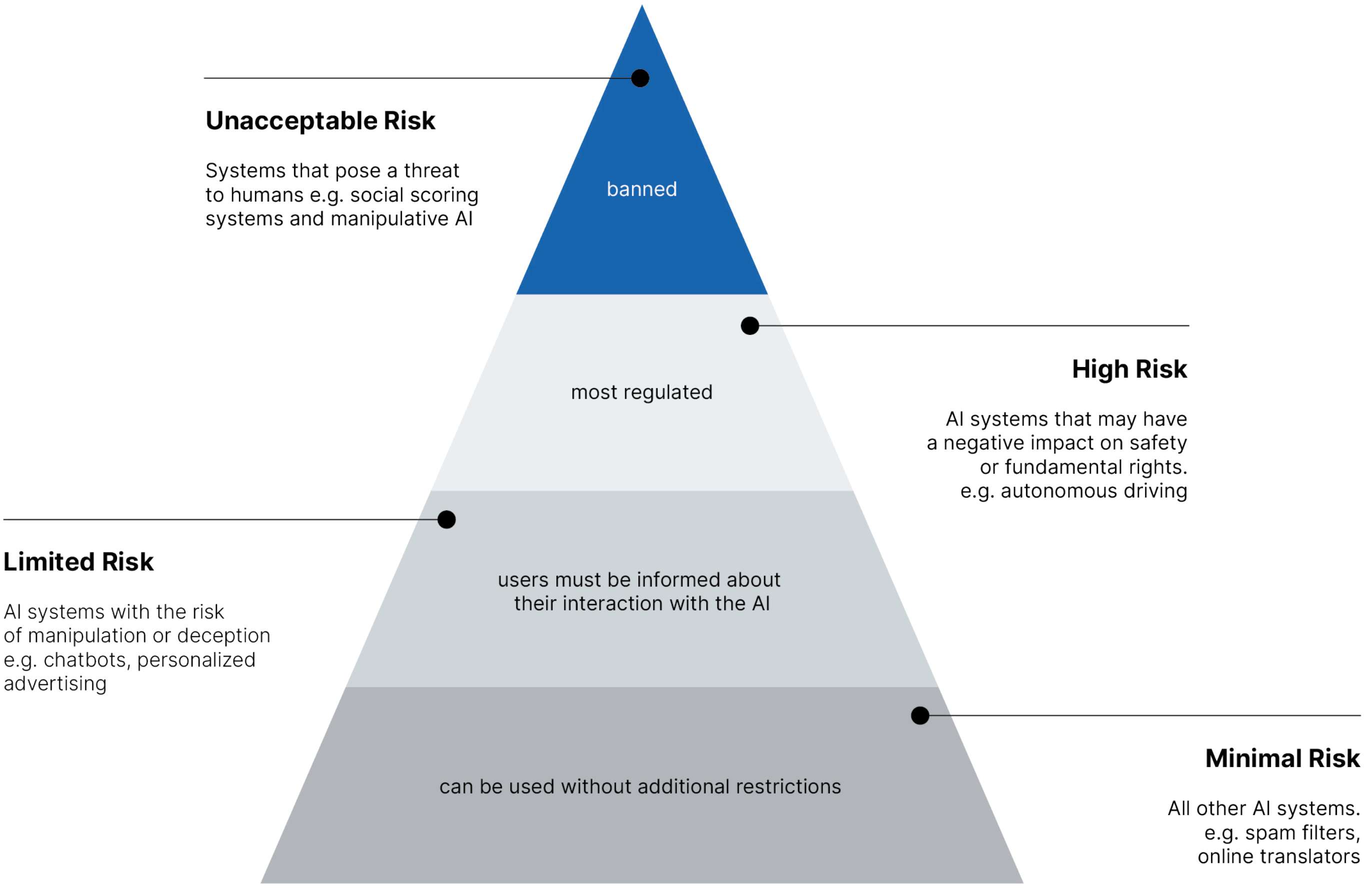

Unacceptable Risk
Systems that pose a threat to humans e.g. social scoring systems and manipulative AI
banned
High Risk
AI systems that may have a negative impact on safety or fundamental rights. e.g. autonomous driving
most regulated
Limited Risk
AI systems with the risk of manipulation or deception e.g. chatbots, personalized advertising
users must be informed about their interaction with the AI
Minimal Risk
All other AI systems. e.g. spam filters, online translators
can be used without additional restrictions

Figure 3: Overview of the EU AI Act's risk classification.

The Four Risk Categories

1. Unacceptable Risk

The highest category, unacceptable risk, includes systems that pose a fundamental threat to individuals. These systems are explicitly listed in article 5 of the AI Act and refer to systems that:

- use **subliminal or deceptive techniques**, impairing informed decision-making and manipulating individuals into making harmful choices they otherwise wouldn't;
- **exploit vulnerabilities** of individuals based on age, disability, or social/economic factors to distort their behavior;
- classify individuals based on their social behavior or characteristics, establishing a **social score**, which can lead to their detrimental or unfavorable treatment;
- **profile individuals** and assess their personality traits to predict their likelihood of committing a crime;
- **create or expand facial recognition databases** through untargeted scraping of internet or CCTV footage;
- **infer emotions of individuals in schools or workplaces**, unless they are employed for medical or safety purposes;

- use **biometric data to infer sensitive personal information**, such as race, political opinions, religious beliefs or sexual orientation;

- use **real-time, remote biometric identification in public places for law enforcement purposes**. Minor exceptions apply here, for instance when such systems are used to locate victims of severe crimes.

All such systems posing an unacceptable risk are **banned** in the EU.

2. High Risk

The next category, high-risk, includes AI systems that may have a negative impact on the safety or fundamental rights of individuals if they fail or are misused. Under article 6 of the EU AI Act, a system is considered high-risk if it is either used in an already regulated product as a safety component (see Annex I) or explicitly designated as high-risk (see Annex III). Products assessed under EU product safety legislation include toys, aircraft and automobiles, medical devices and elevators. For example, a voice-controlled toy that passes on dangerous information to children, or AI applications in robot-assisted surgery. Additionally, AI systems are classified as high-risk if they are used within the following categories:

- critical infrastructure, e.g. in energy supply or digital infrastructure

- education, e.g. assessing students in an exam or determining access to educational institutions

- employment, e.g. in recruiting or making a promotion decisions

- access to essential public and private services, e.g. credit scoring of individuals, or risk assessment in the context of life insurance

- law enforcement, e.g. assessing the personality traits of someone

- migration, e.g. to decide on visa or residence permits

- justice and democratic processes, e.g. to interpret facts or the law

To manage and mitigate the risk such systems may pose, the AI Act lays down certain requirements in articles 9 to 15, that must be fulfilled by the providers of high-risk systems.

Concisely, **providers** must make sure to:

- establish a **continuous risk management system** to oversee their system and ensure compliance throughout its lifecycle. This includes mitigating risks related to intended use as well as **predictable misuse**.

- implement **robust data governance practices** to ensure proper collection, processing, and protection of training and testing data. Data should be **complete, correct** and **relevant** to the model's intended use and sufficiently **unbiased**.

- draw up **thorough technical documentation** covering system design specifications, capabilities, constraints, and regulatory compliance, as well as decisions about how the system was developed to ensure **transparency**.

- **log activity** to ensure **traceability** of operations and results.

- equip deployers of AI systems, with comprehensive information to comply with regulations. This includes **explicit instructions on system usage**, output interpretation, and risk mitigation.

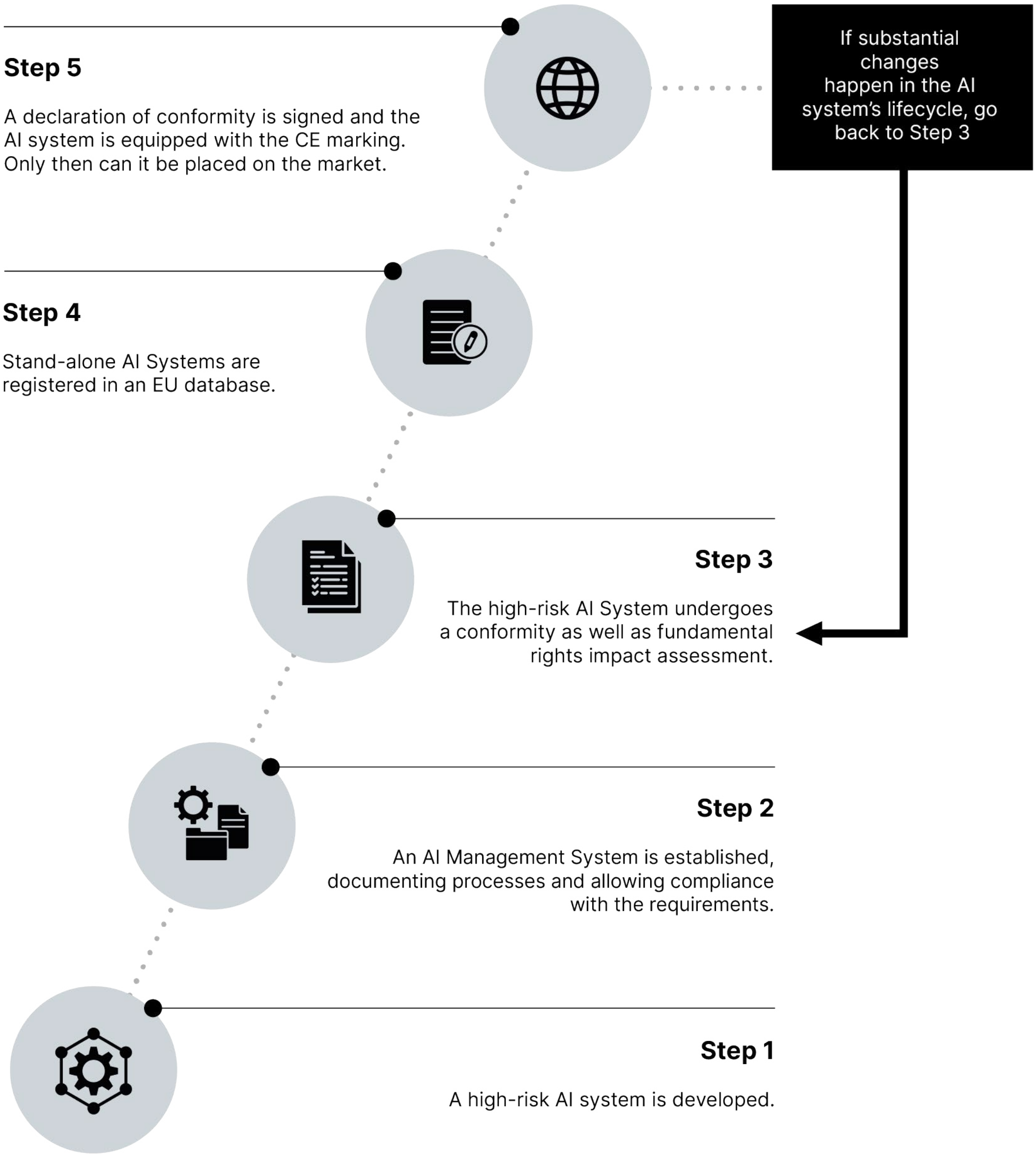
- design systems to **support appropriate human oversight** measures including with **appropriate human-machine interface tools**, enabling users to monitor, override, and intervene with system operations.

- ensure that AI systems maintain **adequate accuracy, robustness,** and **cybersecurity**. This involves establishing backup systems, developing bias-free algorithms, and deploying cybersecurity controls.

To set up an effective risk management system and to ensure that the documentation of the AI systems is sufficient and aligned with the AI Act are likely the most extensive requirements. The requirements above can be summarized in one key obligation of high-risk AI system providers: to establish a "**quality management system**" to control any AI governance and compliance measures. In case an AI system falls into one of the high-risk categories, yet it does not present a substantial threat to health, safety or human rights, providers can evade the requirements mentioned above. It is up to the provider to prove that the system does not pose such a risk, and regulators can sanction organizations for not complying with the requirements.

The steps that must be taken to ensure compliance of high-risk systems can be summarized in Figure 4.

Figure 4: Overview of the steps towards compliance for high-risk AI Systems.



Step 5

A declaration of conformity is signed and the AI system is equipped with the CE marking. Only then can it be placed on the market.

If substantial changes happen in the AI system's lifecycle, go back to Step 3

Step 4

Stand-alone AI Systems are registered in an EU database.

Step 3

The high-risk AI System undergoes a conformity as well as fundamental rights impact assessment.

Step 2

An AI Management System is established, documenting processes and allowing compliance with the requirements.

Step 1

A high-risk AI system is developed.

Deployers of high-risk AI systems, while less restricted than the providers, must also adhere to certain obligations in their use of high-risk AI systems. Roughly summarized, deployers must:

- use the AI system in conformity with the provider's instructions for use,
- appoint qualified individuals to oversee the AI system,
- make sure that input data is relevant and representative,
- monitor the operation of the AI system based on instructions for use and promptly report risks as well as serious incidents to the provider and other relevant actors,
- retain logs automatically generated by the AI system,
- inform workers' representatives and affected workers about the system's use, when deploying high-risk AI systems in the workplace,
- comply with registration requirements (Article 49), and in the case that a system isn't registered in the EU database (Article 71), deployers must refrain from using it and notify the provider or distributor,
- obtain authorization from a judicial or administrative authority when using a high-risk AI system for targeted searches related to criminal offenses, and decisions based solely on system output are prohibited,
- inform individuals that they are subject to the use of a high-risk AI system if it makes decisions related to them,
- cooperate with relevant authorities to implement this regulation.

3. Limited Risk

The last category of AI systems that is regulated by the AI Act are those with a limited risk. Both providers and deployers of AI systems falling under this category must comply with significantly lower requirements, namely certain transparency obligations, to either inform the user about an interaction with an AI or to label AI-generated content.

The AI Act lays down these obligations:

- Providers should **explicitly disclose** to users that they are **interacting with artificial intelligence**. For example, a chatbot should indicate that it is an AI-driven chatbot.
- Deployers of emotion recognition or biometric categorization systems–that are allowed within the scope of the AI Act–must **inform individuals exposed to the system** about its operation. They are to process **personal data** in accordance with the relevant EU regulations.
- Providers of (generative) AI systems that produce synthetic content (audio, image, video, or text) must **label these outputs** as artificially generated or manipulated in a machine-readable format.
- Deployers of an AI system generating **deep fakes** must **label** such content explicitly. Deployers of an AI system that generates or manipulates text on topics of public interest, such as news articles, must indicate this text as AI-generated **unless it has been reviewed and approved by a human editor**.

4. Minimal Risk

All other AI systems, not included in the categories listed above pose minimal to no risk under the AI Act; consider for instance spam filters or AI-powered translators. Systems that fall within this category may be utilized with no further restrictions, although a code of conduct and following transparency requirements are encouraged.

General Purpose AI

For GPAI models (often GenAI), the risk classification depends on the use case of the AI system, which is of course difficult to narrow down in the context of a GPAI – a model that is adaptable and of general usability. Specifically, a GPAI model under the AI Act is defined as an AI model 'that displays significant generality and is capable of competently performing a wide range of distinct tasks regardless of the way the model is placed on the market and that can be integrated into a variety of downstream systems or applications' (Article 3, paragraph 63).

Since it's difficult to regulate the use cases of such models due to their adaptive and flexible nature, the EU AI Act was revised in 2023 to include a distinct set of rules explicitly for GPAI. The regulation of GPAI was arguably the most controversial aspect of the AI Act; it was so heavily debated that negotiations between the EU member states nearly came to a halt[19]. Nonetheless, in December 2023 a consensus on the EU AI Act was reached, establishing a harmonized set of rules for GPAI models. Article 53 of the AI Act specifies which additional rules apply for GPAI models. Note that these obligations only apply to GPAI model providers, not deployers.

Revisiting the exemplary case of the GPAI system ChatGPT, OpenAI is the provider, while any company that makes internal, professional use of ChatGPT, is a deployer. Now, if such a company built an AI application based on GPT-4 to sell on to its customers, then said company would automatically also be a provider – albeit a so-called "downstream provider". OpenAI would then be an upstream provider that would be required to provide the necessary documentation on the model's training. The requirements for providers of GPAI models can be summarized as follows:

- Providers must **maintain up-to-date technical documentation** for their AI models, including details about training, testing, evaluation, as well as known or **estimated energy consumption** of the model. This information should be made **available to the AI Office**, a new responsible authority created by the European Commission, and other relevant national authorities upon request.

- Providers must **share relevant information** and documentation with other AI system providers who aim to integrate the general-purpose AI model into their systems, such that they can fully comprehend the model's **capabilities** and **limitations**, and hence ensure its compliance.

- Providers must **establish a policy** to comply with Union law on copyright provisions and related rights.

- Providers must create a **detailed summary of the content** used for **training** the general-purpose AI model, in accordance with a **template** that is to be provided by the AI Office.

- Providers **outside the EU** that want to place their GPAI model on the EU market need to **appoint an authorized representative** who is established in the EU and who performs the obligation tasks under the AI Act.

At time of writing, providers of GPAI AI models released under free and open-source licenses and which meet certain conditions are exempt from most obligations here. For instance, providers of such models need not maintain a technical documentation record, or share information with downstream providers, if their model's licenses allow for the access, usage, modification, and distribution of the model, as well as insight into the model's exact parameters.

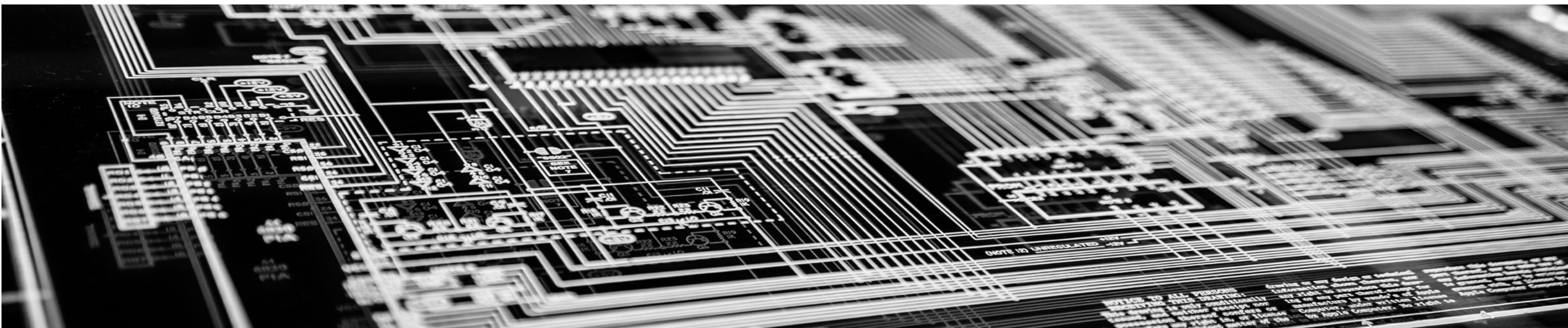
Further, a separate class of GPAI models has been defined: **GPAI models posing a systemic risk**. A GPAI is classified as having a systemic risk, if it has been trained with a particularly high processing power (10^25 floating point operations) or similar. Such models–whether open source or not–are subject to stricter requirements, which include access, usage, modification, and distribution of the model, as well as insight into the model's exact parameters. Such models (whether open source or not) are subject to stricter requirements (Article 55 of the AI Act), which include:

- Immediately notifying the European Commission about the GPAI model without delay
- Model evaluation and attack tests for risk assessment and mitigation
- Tracking and reporting of serious incidents
- Cybersecurity protection measures
- Additional documentation requirements

The enforcement and supervision of GPAI models is to be carried out by the AI Office with the support of a competent scientific committee that is to be formed. Additionally, each member state is responsible for setting up its own national authority to oversee matters related to the AI Act.

Conclusion and Consequences

In conclusion, the EU AI Act is the world's first comprehensive AI law. Striking a balance between AI innovations and their immense potential, and regulating the significant risks they bring about, is a challenging endeavor and only time will tell whether this first attempt is a step in the right direction, namely a step towards responsible and trustworthy AI. Many organizations, especially those located in Europe, will now need to prepare for the AI Act, which creates both additional compliance overhead and costs. However, not meeting the requirements of the AI Act will be even more costly. **Non-compliance** with certain AI practices can result in **fines up to 35 million EUR** or **7 % of a company's annual turnover** (e.g. when placing a prohibited system on the market). Other violations can result in **fines up to 15 million EUR** or **3 % of a company's annual turnover** (e.g. for non-compliance with high-risk requirements). Even **providing incorrect or misleading information** can result in **fines up to 7.5 million EUR or 1 % of a company's annual turnover**. SMEs will receive lower fines.



Criticism

While the necessity of such a regulation has become clear through numerous AI related scandals and lawsuits, such as the NYT lawsuit against OpenAI[20], not everyone is convinced of the effectiveness of the Act. To be precise, concerns have been raised about the impact of these new regulations on innovation and competitiveness, as well as about the lack of clarity on certain matters[21]. Let us discuss these two points in more detail.

First, many expressed the fear that the AI Act will create additional regulatory barriers that will favor American and Chinese competition, limiting the opportunities of European AI champions. While the EU is taking an anthropocentric approach, centered around consumer protection, fairness and safety, the US has mostly focused on self-regulation and innovation promotion[22] – notably though, they have recognized the necessity of stricter regulation and started following suit to the EU AI Act. In this respect, the Biden Administration issued the 'Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence'[23], while certain regulations have also been proposed. A prominent example is the Colorado Bill[24], some parts of which strongly resemble the EU AI Act, and which is the first U.S. law expressly regulating the use of AI. Prior to that, the New York City law on AI bias[25] was enforced in 2023, making it unlawful for employers to utilize automated employment decision tools for candidate or employee screening, unless specific bias audit and notice requirements are met. Meanwhile, China's approach has been characterized as centered around state control and economic dynamism[22].

To examine the potential impact of the EU AI Act on the European AI ecosystem, an early study was conducted by AppliedAI in 2022[26], questioning more than 100 AI Startups and other companies developing AI across Europe, complemented by the views of 15 Venture Capital Firms. The results, as illustrated in Figure 5, showed that about half of the AI Startups believe the AI Act will slow down AI Innovation in Europe, while roughly a sixth considers halting the development of AI or relocating outside the EU.

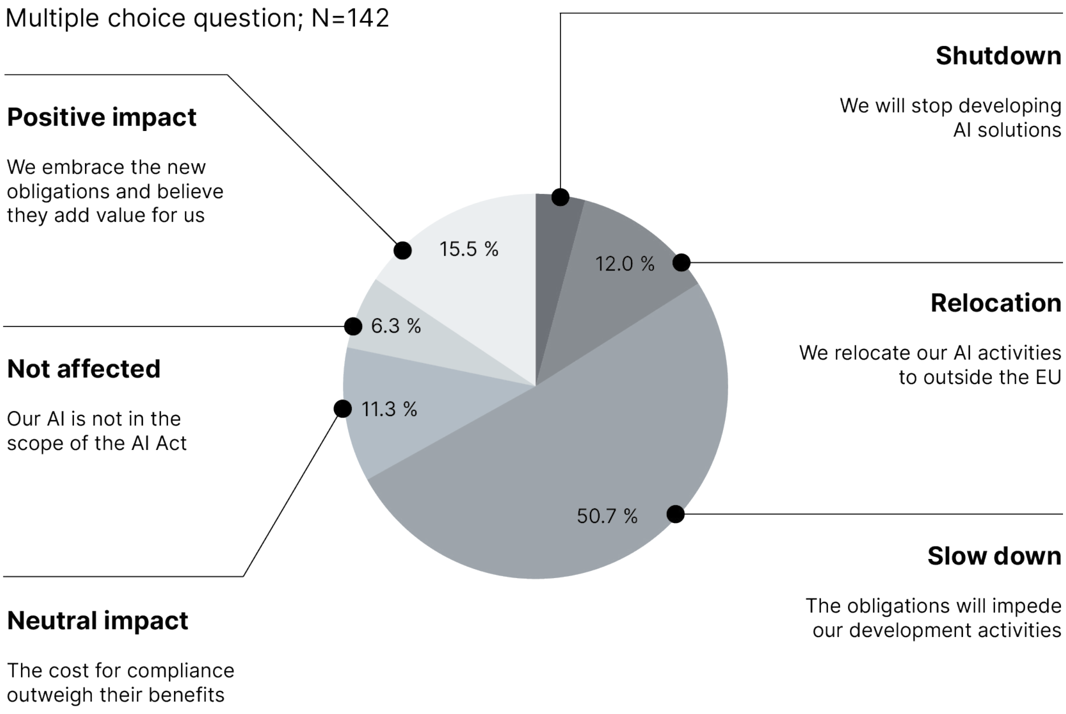


Figure 5: Overview of the estimated impact the AI Act will have on innovation and competitiveness, as assessed by startups in 2022. Two thirds of the surveyed startups expected a negative impact of the AI Act on Innovation in Europe.

Reprinted and adapted with permission from 'AI Act Impact Survey' by AppliedAI, 2022 [26].

Concurrently, VCs also expressed concerns about Europe's competitiveness in the field of AI, considering divesting, as depicted in Figure 6.

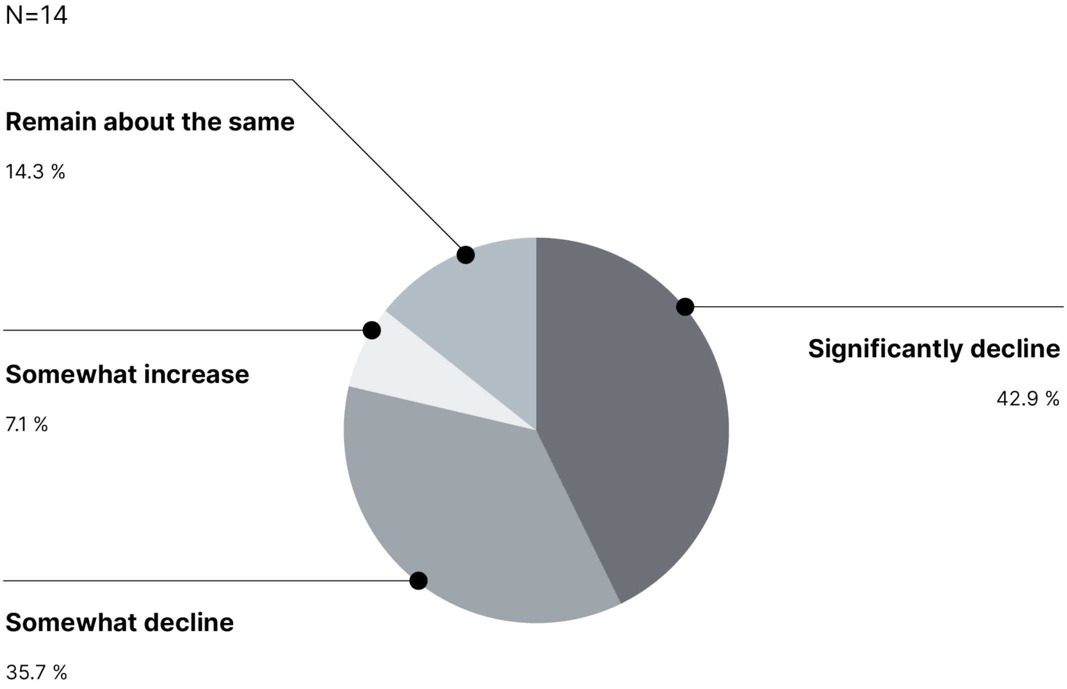


Figure 6: Overview of the estimated impact the AI Act will have on innovation and competitiveness, as assessed by VCs in 2022. VCs foresee a drastic decline in Europe's competitiveness in AI and will shift their investments either outside of Europe or into other technologies.

Reprinted and adapted with permission from 'AI Act Impact Survey' by AppliedAI, 2022 [26].

The German AI Federal Association (KI Bundesverband) also conducted a similar study in 2021[27], “Startups and artificial intelligence - innovation meets responsibility” (Startups und Künstliche Intelligenz – Innovation trifft Verantwortung). The study highlights the fact that the USA and China currently lead in AI technology, while Europe risks falling behind – raising concerns about digital sovereignty and the need to promote AI innovations.

At the same time, there is a basic consensus among the surveyed German AI start-ups on the need for ethical guidelines; 88 % want to assume social responsibility and 81 % believe that ethical issues must also be taken into account when developing the technology. Almost half of the surveyed startups view European AI regulation positively in terms of creating trust and a European unique selling point. Nonetheless, regulation brings about legal uncertainties and questions of practical feasibility, which pose major challenges for AI start-ups. Undoubtedly, European regulation must balance fostering trust through responsible AI with preventing repercussions on innovation, especially for startups.

Advancing AI education, promoting developer diversity, and fostering collaboration between startups and established industries are suggested as necessary steps to ensure competitiveness by translating Europe’s research strengths into economic applications, and hence, driving European AI forward.

To address concerns on competitiveness, the European Commission revised the AI Act to include measures in support of innovation. Specifically, Article 57 introduces AI regulatory sandboxes, premised to provide controlled environments for start-ups and small and medium-sized enterprises to develop, train, and test novel AI systems prior to their market launch. Risks related to fundamental rights, health, safety, and testing are to be assessed within the sandboxes. Additionally, competent authorities are to offer guidance to providers participating in the AI regulatory sandbox on matters of regulatory expectations and achievement of compliance with the Act. In this way, an attempt at a legal pathway for safe experimentation was made, which simultaneously aims at facilitating innovation and fostering the development of an AI ecosystem.

Concurrently, setting clear boundaries is expected to build user trust, which in turn leads to increased demand and development of AI systems. Regulatory sandboxes must be operational within 24 months of the enactment of the regulation. To elucidate whether the European Commission’s innovation measures are in fact perceived as effective, studies such as the ones mentioned above, which were conducted by AppliedAI and the German AI Federal Association in the years 2022 and 2021 respectively, could be repeated, to examine whether the general consensus among start-ups on the impact of the AI Act has since improved.

An additional point of criticism has been the uncertainty and complexity of classification. To be precise, classifying AI systems into the categories described above is not always unambiguous. Results from a study AppliedAI conducted on ‘Risk Classification of AI systems from a Practical Perspective’ in 2023[28] yielded a high proportion of unclear risk assessments across various domains, reaching almost 40 % of the examined cases. This uncertainty can be quite problematic, as it may result in companies hesitating to invest in AI and adopt AI technologies due to fear of errors or penalties. ‘Legal hurdles’ are additionally cited as an obstacle to AI deployment by almost half of the surveyed companies. Addressing these uncertainties is crucial to enable enterprises to fully harness the potential of AI.

Since the territory the AI Act explores is still new and uncharted, we expect these uncertainties to be clarified step by step in the coming months. Standards, templates, guidance documents, classification instructions and examples are necessary to minimize bureaucracy and support companies in tackling the current ambiguities. Offering assistance with proper risk classification may be necessary to achieve clear and assured decisions and to truly drive the impact of responsible AI.

Take Action!

All in all, reducing current uncertainties is vital, as it will enable companies to confidently pursue AI-driven innovations, marking the beginning of a new chapter in the history of AI, where innovation is intertwined with responsibility, safety, and ethics. And in the future of AI, it's not just about what technology can achieve, but about doing it right.

Strategic consulting offers guidance in finding the right tech-stack and building compliant solutions. Staying up-to-date with current regulations, while thinking long-term, can be difficult. Munich-based KI-Lab has been a trusted advisor to top management of large and medium-sized companies across all sectors. Our collaboration with the Technical University of Munich ensures we continuously enhance our consulting tools, delivering maximum value in various areas, including Data Science & AI Applications. As the EU AI Act brings about new compliance challenges, let the KI-Lab guide you through them.

We can provide you with clarity on relevant issues of AI regulation, data protection as well as best practices.

Hereby it is crucial to take an innovative approach to AI compliance and governance, to minimize the manual overhead and facilitate efficient as well as trustworthy AI deployment. Our strategic partner trail, in turn, can assist you with the technological know-how. trail developed a software solution to help companies of all sizes efficiently comply with AI regulation. Their AI Governance Copilot automates and orchestrates regulatory requirements throughout the AI lifecycle and enables quick and easy audits and certification of AI systems at minimal overhead for data scientists.

Take action now – visit https://ki-lab.net/ to learn more about the KI-Lab and discover trail's AI regulatory compliance solution at www.trail-ml.com.

References

1. Yagoda, M. (2024, February 23). Airline held liable for its chatbot giving passenger bad advice – what this means for travellers. BBC. https://www.bbc.com/travel/article/20240222-air-canada-chatbot-misinformation-what-travellers-should-know

2. Barshay, J. (2024, July 8). PROOF POINTS: Asian American students lose more points in an AI essay grading study – but researchers don’t know why. The Hechinger Report. https://hechingerreport.org/proof-points-asian-american-ai-bias/

3. Kelly, J. (2024, May 31).Google’s AI Recommended Adding Glue To Pizza And Other Misinformation – What Caused The Viral Blunders? Forbes. https://www.forbes.com/sites/jackkelly/2024/05/31/google-ai-glue-to-pizza-viral-blunders/

4. Héder, M. (2020). A criticism of AI ethics guidelines. Információs Társadalom, 20(4), 57. https://doi.org/10.22503/inftars.XX.2020.4.5

5. Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., & Song, D. (2021). Natural adversarial examples. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 15262-15271). https://arxiv.org/pdf/1907.07174

6. Xu, H., Ma, Y., Liu, H.-C., Deb, D., Liu, H., Tang, J.-L., & Jain, A. K. (2020). Adversarial Attacks and Defenses in Images, Graphs and Text: A Review. International Journal of Automation and Computing, 17(2), 151–178. https://doi.org/10.1007/s11633-019-1211-x

7. Díaz-Rodríguez, N., Del Ser, J., Coeckelbergh, M., López de Prado, M., Herrera-Viedma, E., & Herrera, F. (2023). Connecting the dots in trustworthy Artificial Intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation. Information Fusion, 99, 101896. https://doi.org/10.1016/j.inffus.2023.101896

8. Corrêa, N. K., Galvão, C., Santos, J. W., Del Pino, C., Pinto, E. P., Barbosa, C., Massmann, D., Mambrini, R., Galvão, L., Terem, E., & Oliveira, N. de (2023). Worldwide AI ethics: A review of 200 guidelines and recommendations for AI governance. Patterns, 4(10), 100857. https://doi.org/10.1016/j.patter.2023.100857

9. OECD AI Policy Observatory Portal. (n.d.). https://oecd.ai/en/ai-principles

10. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. (2018). Ethically Aligned Design, Version 2 – Overview. In Ethically Aligned Design, Version 2 – Overview [Report]. https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_brochure_v2.pdf

11. High-Level Expert Group on Artificial Intelligence. (2019). A definition of AI: main capabilities and scientific disciplines. In European Commission. https://www.aepd.es/sites/default/files/2019-12/ai-definition.pdf

12. Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

13. European Commission. (2022, June 7). High-level expert group on artificial intelligence: Shaping Europe’s digital future. Digital Strategy. Retrieved August 9, 2024, from https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai

14. European AI Alliance – Welcome to the ALTAI portal! (n.d.). Futurium. https://futurium.ec.europa.eu/en/european-ai-alliance/pages/welcome-altai-portal

15. University of Maryland. (n.d.). Institute for Trustworthy AI in Law & Society (TRAILS). https://www.trails.umd.edu/

16. Tabassi, E. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). https://doi.org/10.6028/nist.ai.100-1

17. Mammonas, D. (2023, December 9). Artificial intelligence act: Council and Parliament strike a deal on the first rules for AI in the world. European Council. https://www.consilium.europa.eu/en/press/press-releases/2023/12/09/artificial-intelligence-act-council-and-parliament-strike-a-deal-on-the-first-worldwide-rules-for-ai/

18. Regulation (EU) 2024/1689 of the European Parliament and of the Council 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)

19. Piquard, A. (2023, November 18). European regulation on artificial intelligence threatened by stalled negotiations. Le Monde. European regulation on artificial intelligence threatened by stalled negotiations (lemonde.fr)

20. Grynbaum, M.M. & Mac, R. (2023, December 27). The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work. The New York Times. https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html

21. Pieper, F-U. & Schmalenberger, A. (2024, February 19). Wie viel Rechtsunsicherheit steckt in Europas KI-Verordnung? TaylorWessing. https://www.taylorwessing.com/de/insights-and-events/insights/2024/02/wie-viel-rechtsunsicherheit-steckt-in-europas-ki-verordnung

22. Trustible (2023, July 18). How Does China's Approach To AI Regulation Differ From The US And EU? Forbes. https://www.forbes.com/sites/forbeseq/2023/07/18/how-does-chinas-approach-to-ai-regulation-differ-from-the-us-and-eu/

23. Kang, C. & Sanger, D.E. (2023, October 30). Biden Issues Executive Order to Create A.I. Safeguards. The New York Times https://www.nytimes.com/2023/10/30/us/politics/biden-ai-regulation.html

24. Colorado General Assembly. (2024). Senate Bill 24-205: Consumer Protections for Artificial Intelligence. 2024 Regular Session. https://leg.colorado.gov/sites/default/files/2024a_205_signed.pdf

25. Masling, S.P., Lee, W.J., Kadish, D.A. & Corcoran, C.M. (2023, April 19). New York City issues final rule on AI Bias Law and postpones enforcement to July 2023. Morgan Lewis. https://www.morganlewis.com/pubs/2023/04/new-york-city-issues-final-rule-on-ai-bias-law-and-postpones-enforcement-to-july-2023

26. Liebl, A. & Klein, T. (2022, December). AI Act Impact Survey. AppliedAI. AI Act Impact Survey_clean slides (appliedai.de)

27. Hirschfeld, A., Gilde, J., Walk, V., Cann, V. & Seitz, J. (2021). Startups und Künstliche Intelligenz – Innovation trifft Verantwortung. KI Bundesverband. https://ki-verband.de/wp-content/uploads/2021/12/dl-studie-startups-ki.pdf

28. Liebl, A. & Klein, T. (2023, March). AI Act: Risk Classification of AI Systems from a Practical Perspective. AppliedAI. https://www.appliedai.de/assets/files/AI-Act-Risk-Classification-Study-EN.pdf

About us

About Trail

trail is a start-up based in **Munich** enabling companies to build trustworthy, high-quality and compliant AI solutions by **automating governance**, and was named as one of the most promising AI start-ups in Germany by appliedAI. The **trail AI Governance Copilot** supports developers and compliance teams in managing AI systems, and in aligning them with internal policies, standards and regulation. The platform makes it easy to operationalize the principles of trustworthy AI while saving time through trail's automation capabilities.

trail

About KI-Lab

The KI-Lab is a joint initiative by the **consultancy company TCW**, renowned for advising top management across diverse industries, and the **Technical University of Munich**, a leading entrepreneurial institution with outstanding scientific and technological expertise. Building on this ecosystem, the KI-Lab bridges the g**ap between corporate needs and technological advancements**, offering pragmatic consulting services powered by data science.

At the KI-Lab, we:

- Develop innovative business models and address strategic and technical challenges regarding data analytics and AI in our workshops.
- Transform creative ideas into projects tailored to your company through our sprint projects.
- Conduct research projects exploring artificial intelligence and data science in both technical and business contexts.
- Leverage the skills of our talented student pool through final theses and data challenges.

Our comprehensive approach ensures that we provide practical, cutting-edge solutions to meet the evolving demands of today's businesses. With the EU AI Act introducing new compliance challenges, we are here to guide you through them. We offer clarity on AI regulation, data protection, and best practices, helping you navigate the complexities of this new regulatory landscape.

KILAB

2024

Publisher:
Univ.-Prof. Dr. Dr. h. c. mult.
Horst Wildemann

ISBN: 978-3-947730-47-6

Authors:
Niklas Stepanek [1]
Ilia Frantzi [1]
Charlotte Schöllkopf [1][2]
Nick Malter [2]
Anna Spitznagel [2]

Design:
Luisa Pantow

Editor:
Univ.-Prof. Dr. Dr. h. c. mult.
Horst Wildemann[1]

Affiliations:
[1] KI-Lab
[2] trail