

# Artificial Intelligence Act:

A Policy Prototyping Experiment

*Human Oversight, Transparency and Risk Management requirements in the AI Act*





## About Open Loop

Open Loop is a global program that connects policymakers and technology companies to help develop effective and evidence-based policies around AI and other emerging technologies. The program, supported by Meta builds on the collaboration and contributions of a consortium composed of regulators, governments, tech businesses, academics, and civil society representatives. Through experimental governance methods, Open Loop members co-create policy prototypes and test new and different approaches to laws and regulations before they are enacted, improving the quality of rulemaking processes in the field of tech policy.

This report presents the findings and recommendations of the 2nd part of the 1st pillar of the program - Operationalizing the Requirements for AI Systems, which was rolled out in Europe from July 2022 to August 2022 in partnership with Estonia's Ministries of Economic Affairs and Communications and Justice and the Malta Digital Innovation Authority (MDIA).

This work is licensed under a [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/).



## Cite this report

Andrade, Norberto Nuno Gomes de, Laura Galindo and Antonella Zarra. "Artificial Intelligence Act: A Policy Prototyping Experiment: Human Oversight, Transparency and Risk Management Requirements in the AI Act" (2023), at [https://openloop.org/wp-content/uploads/2023/06/AI\\_Act\\_Human\\_Oversight\\_Transparency\\_Risk\\_Management\\_requirements.pdf](https://openloop.org/wp-content/uploads/2023/06/AI_Act_Human_Oversight_Transparency_Risk_Management_requirements.pdf)

## Acknowledgements

This policy prototyping program was co-designed and facilitated by Meta in collaboration with our partners from the Government of Estonia and the Malta Digital Innovation Authority (MDIA). We want to thank in particular Henrik Trasberg, Legal Advisor on AI and New Technologies, Estonian Ministry of Justice; Ott Velsberg, Government Chief Data Officer, Estonian Ministry of Economic Affairs and Communications; Kenneth Brincat, CEO, MDIA and Ian Gauci and Gordon Pace, Advisors, MDIA. A special thanks to Prof. Bart Schermer and Jord Goudsmit from Considerati for their contribution to the project.

We would like to thank all the individuals who provided feedback on earlier versions of this report, particularly to Fuz Dudhwala, Nick Manzoli, Svetlana Matt, and Maeve Ryan for their insightful reviews of the draft.

We are particularly thankful to all the speakers and participants of the workshop held on 14 November 2022 for informing this report through their excellent contributions:

Victoria Brugada Ramentol - Virtuleap, Kai Zenner - European Parliament, Maarten Stolk - Deeploy, Risto Uuk - Future of Life Institute, Clara Neppel - IEEE, Aram Markosyan - Meta, Alexandru Adrian Tantar - Luxembourg Institute of Science and Technology, Dan Adamson - Armilla AI, Johann Laux - Oxford Internet Institute, Jan-Kees Buenen - Synerscope, Björn Preuss - 2021.ai, Katya Klinova - Partnership on AI.









## AI Providers

Thank you in particular to the individual experts that represented the participating companies.

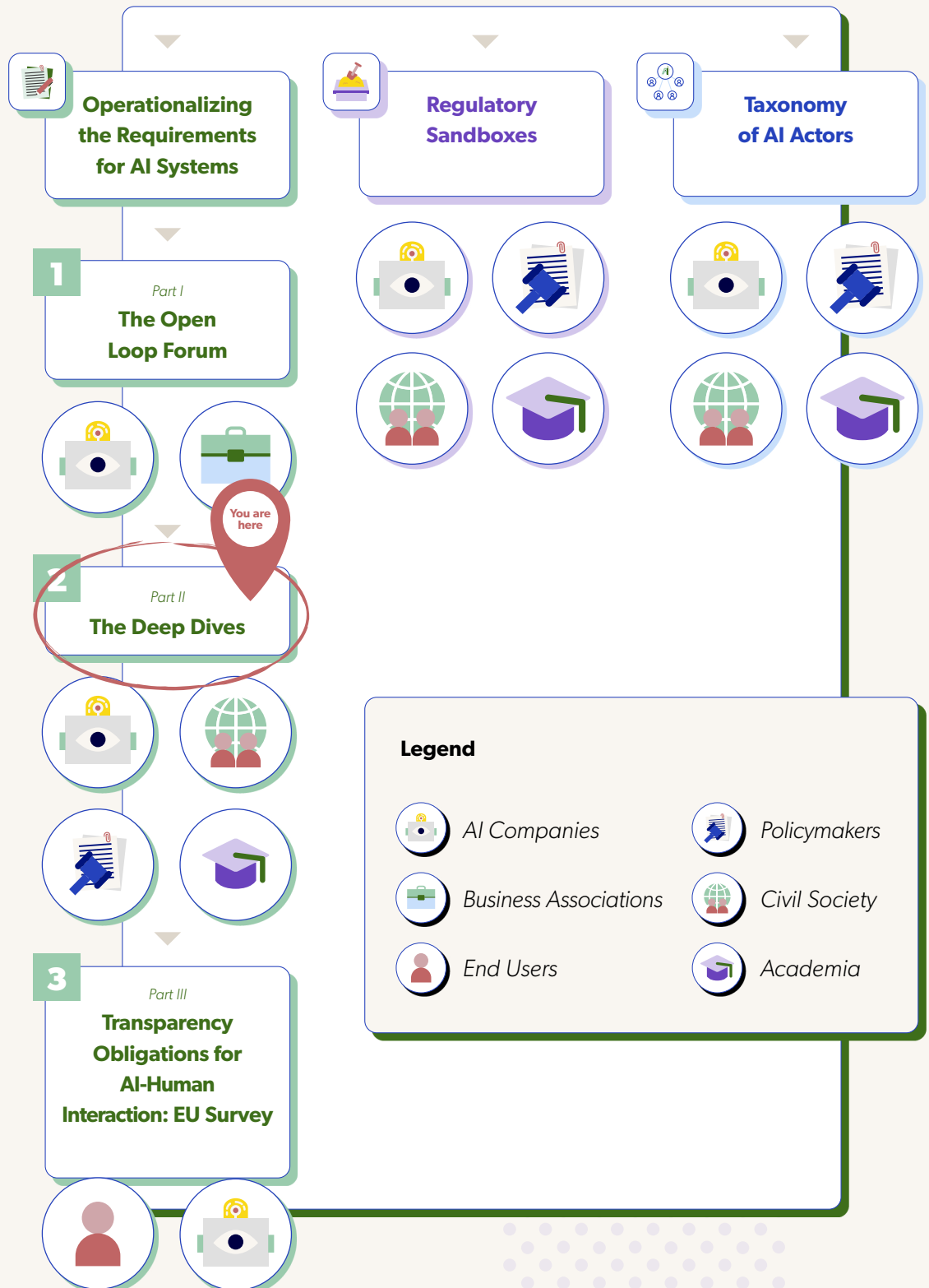
Company	Expert
Keyless	Pierluigi Failla
LearnerShape	Maury Shenk
Peregrine Technologies mbH	Philip Meier
Virtuleap	Amir Bozorgzadeh Victòria Brugada-Ramentol Bebiana Moura
SynerScope bv	Jan-Kees Buenen
The Newsroom	Pedro Henriques Jenny Romano Lorenzo Mora
ASC27	Colum Donnelly Nicola Grandis
DLabs	Maciej Karpicz
Telesoftas	Bart Kappel
Vedrai SPA	Hossein (Kian) Sarpanah

Advisors

Company	Expert
 2021.ai	Jesper Krognos
 CredoAI	Ehrik Aldana Susanna Shattuck Evi Fuelle
 DEEPLY Deeploy	Bastiaan van de Rakt Maarten Stolk
 Enzai Technologies Limited	Ryan Donnelly
 Zupervise	Janhvi Pradhan Deshmukh
 Armillia AI	Philip Dawson

# Artificial Intelligence Act: A Policy Prototyping Experiment

Overview of the Open Loop Program on the AI Act and the stakeholders involved



<b>Executive summary</b>	<b>8</b>
--------------------------	----------

<b>1 Introduction</b>	<b>11</b>
The bigger picture .....	12
Program goals .....	13

<b>2 Methodology</b>	<b>14</b>
Theory of Change for transparency and human oversight requirements .....	15
Theory of Change for risk requirements .....	18
Program Description: A Policy Prototyping Experiment on the AIA .....	19
Participants .....	20
Data Collection .....	21
Limitations .....	22

<b>3 Part 1: Transparency</b>	<b>24</b>
Introduction .....	25
Activity 1: Human Oversight (Article 14) in the AIA .....	25
Activity 2: Transparency (Article 13) .....	29
Activities 3 & 4: The benefits of technical guidance .....	32

**4 Part 2: Risk management in the AIA 37**

Introduction .....	38
Activity 1: Describing the development of the AI system .....	38
Activity 2: Pre-determined changes.....	41
Activity 3: Monitoring.....	43
Activity 4: Validation and testing.....	45

**5 Conclusion and recommendations 47**

**Endnotes 51**



# Executive summary





## Executive summary

This report, which is part of the Open Loop Program on the EU Artificial Intelligence Act (AIA), presents the findings of a policy prototyping exercise on risk management and transparency in the AIA. The objective of this Deep Dive was to assess the clarity and feasibility of selected requirements from the perspective of participating AI companies. Valuable insights gathered from participating companies have highlighted areas that require improvement, which are discussed in two parts: i) transparency and human oversight and ii) risk management requirements in the AIA.

### Main insights of the report



#### **Transparency and Human Oversight Requirements**

The transparency requirement (Article 13) of the AIA does not adequately consider diverse audiences needing human oversight and interpretability of AI systems' outputs. For instance, certain errors, like model drift, may be challenging or impossible to detect for users without assistance from the system provider.

The requirement of human oversight (Article 14) lacks efforts in centralizing, documenting, and making methods available to the public. Additional guidance and standardization are necessary, along with clarification on the division of responsibility for human oversight to enhance feasibility.

Participants emphasized that technical guidance would be more useful than abstract legal requirements, considering the complexity of the AIA.



#### **Risk Management Requirements**

Describing and documenting the development of AI systems (Article 11(1) and Annex IV) poses challenges for participants due to concerns about trade secrets and sensitive information. The administrative and compliance burden associated with this requirement is significant.

Furthermore, the terms "substantial modifications" and "pre-determined changes" lack clarity in the context of AI systems, making them unfeasible.

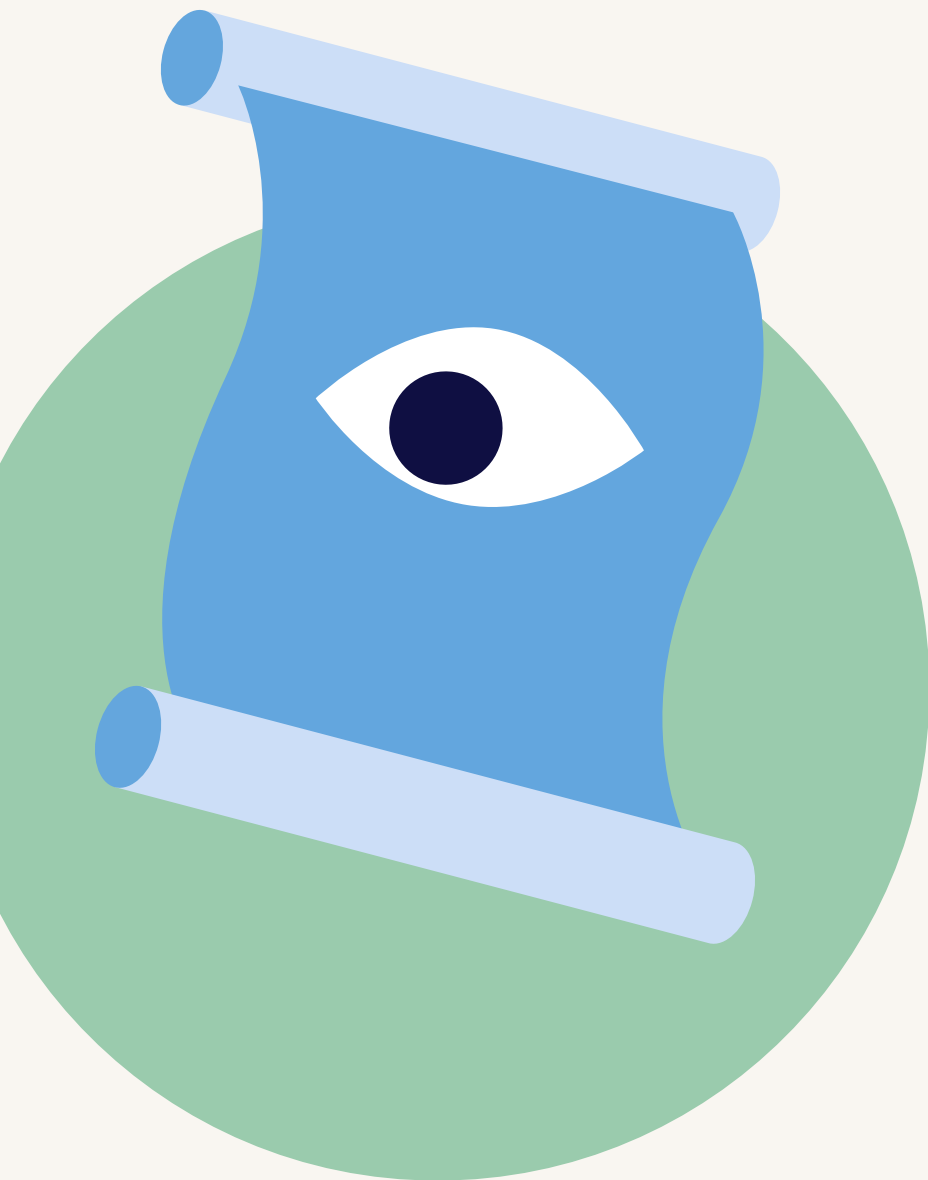
Responsibility for monitoring the operation of AI systems depends on how they are placed on the market, necessitating greater clarity in the division of responsibility.

Additionally, the report reveals significant variation among participants in measuring accuracy, robustness, cybersecurity, and discriminatory impact of AI systems.

Clear and uniform metrics and standards are essential to enable consistent and objective evaluation of AI systems across different use cases and contexts.

**Recommendations  
to policymakers**

- Regarding Transparency (Article 13), explore a 'modular approach' to provide instructions that enable a hands-on approach while allowing for standardization.
- For documenting the development of AI systems, consider exempting third-party models from the requirement or, alternatively, require third-party providers to issue technical documentation to supervisory authorities and/or clients for compliance purposes.
- Provide further clarification on what constitutes a substantial modification in the context of AI and how it should be measured and defined.



1

# Introduction



This report presents the findings of the 2nd part of the 1st pillar of the program - "**Operationalizing the Requirements for AI Systems.**" The objective was to examine the clarity, feasibility, and costs of the transparency and risk management requirements outlined in the EU AI Act (AIA).

## The bigger picture

The first part of the policy prototyping program involved gathering input from over 50 companies subject to the AIA requirements. Their feedback, obtained through multiple-choice and open questions, was published in November 2022.<sup>1</sup> This report focuses on the 2nd part of the 1st pillar of the program, which involved conducting in-depth assessments, or "deep dives," focused on the AIA requirements regarding transparency, human oversight and risk management.

In addition to this report, the program includes a third part that examines transparency obligations for AI-human interaction, specifically studying individuals' responses to notifications and the impact of different information designs.



## Program goals

This specific part of the program was guided by the following overarching goals:

- 1 Gain insight into the clarity, actionability, and comprehensibility of selected AIA requirements.
- 2 Understand the policy implementation challenges, costs, and technical feasibility associated with these requirements.
- 3 Determine whether more concrete guidance, such as a playbook on AI transparency, would benefit stakeholders.<sup>2</sup>



### The specific goals regarding the assessment of the transparency requirements' included:

- Understanding the measures implemented by providers to enable human oversight.
- Assessing the technical measures in place for interpreting AI system outputs (transparency).
- Evaluating the need for a playbook on AI transparency to operationalize AIA requirements in Article 13.
- Determining the level of prescriptiveness required for transparency requirements and the potential benefits of a playbook.



### The specific goals regarding the assessment of the risk management requirements included:

- Examining how users describe their AI systems and assessing its sufficiency for compliance determination and associated administrative costs.
- Testing the clarity and feasibility of the concept of pre-determined changes and its definition.
- Assessing participants' ability to implement technical monitoring of AI systems after deployment.
- Ensuring user understanding of the "metrics used to measure accuracy, robustness, cybersecurity, and compliance with Chapter II."

By addressing these goals, this program aimed to provide policymakers with valuable insights to enhance the implementation and effectiveness of the AIA's transparency and risk management provisions.

# 2

## Methodology



In a policy prototyping exercise, we assess if the policy prototype is effective in what it wants to accomplish. In other words, we assess whether the requirements of the prototype (in this case the proposed AIA) lead to the desired outcomes which contribute to the (long term) goals of the policy.

To assess the effectiveness, we employ the Theory of Change (ToC) approach (Figure 1). The ToC illustrates how the requirements should result in the desired outcomes and contribute to the policy goals. We begin by defining the goals of the requirements and then test their clarity, actionability, and comprehensibility for the intended recipients. Additionally, we evaluate the implementation of the policy in terms of cost and technical feasibility. By understanding the ability of the recipients to fulfill the requirements, we can assess whether the desired outcomes are achieved and contribute to the policy sub-goals.

Here’s a simplified example of a Theory of Change:

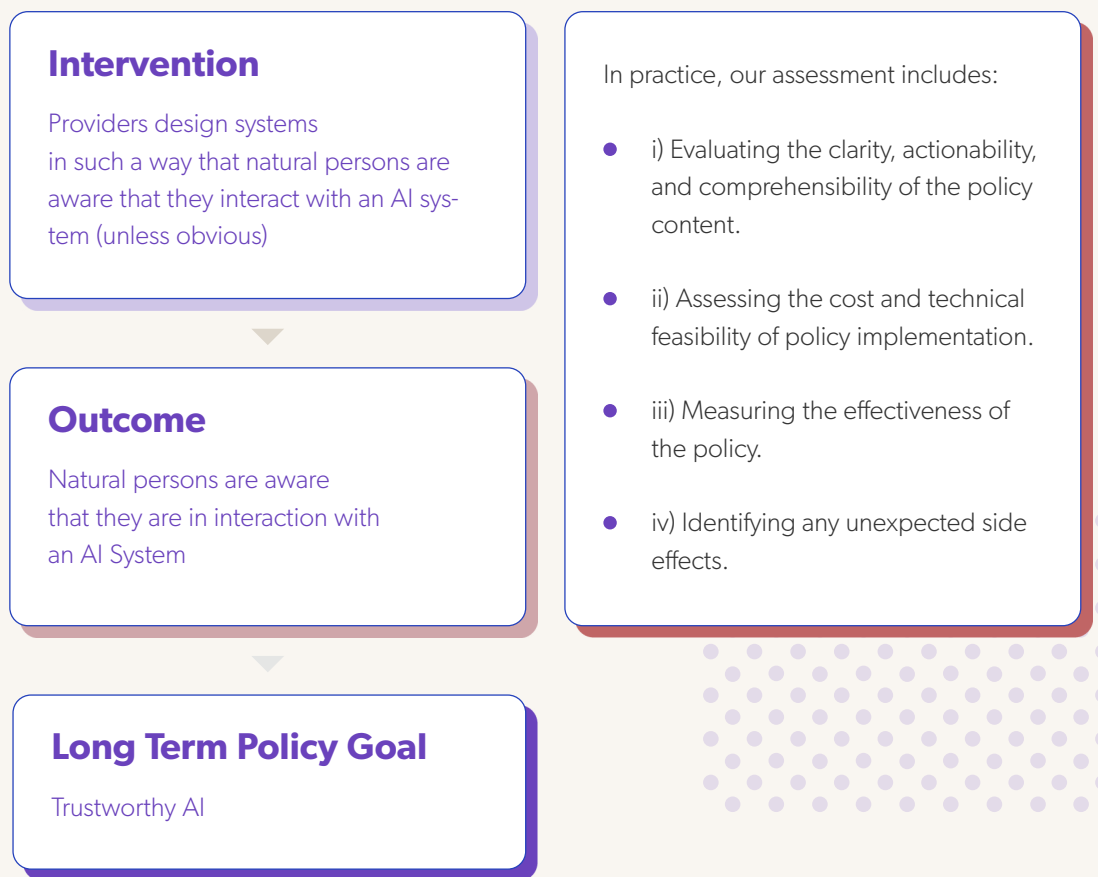


Figure 1. Example of a Theory of Change



### Theory of Change for transparency and human oversight requirements

The long-term policy goal is to have trustworthy AI, which effectively mitigates risks to health, safety, and fundamental rights. From the AIA, we derive two policy sub-goals related to transparency (Figure 2).

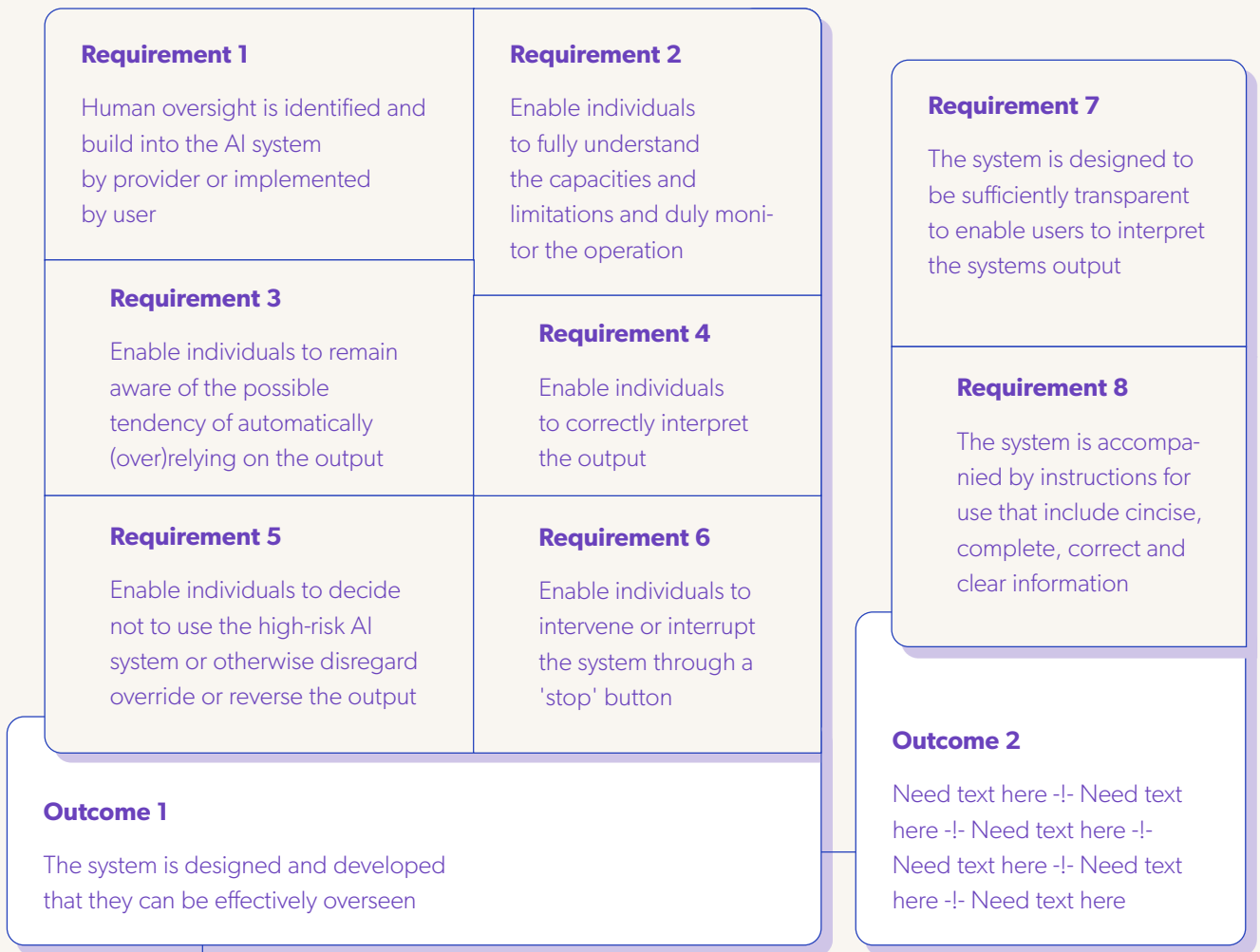


Figure 2. Theory of change for transparency requirements in the AIA



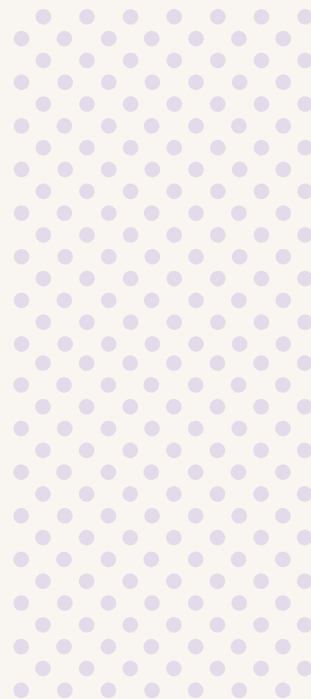
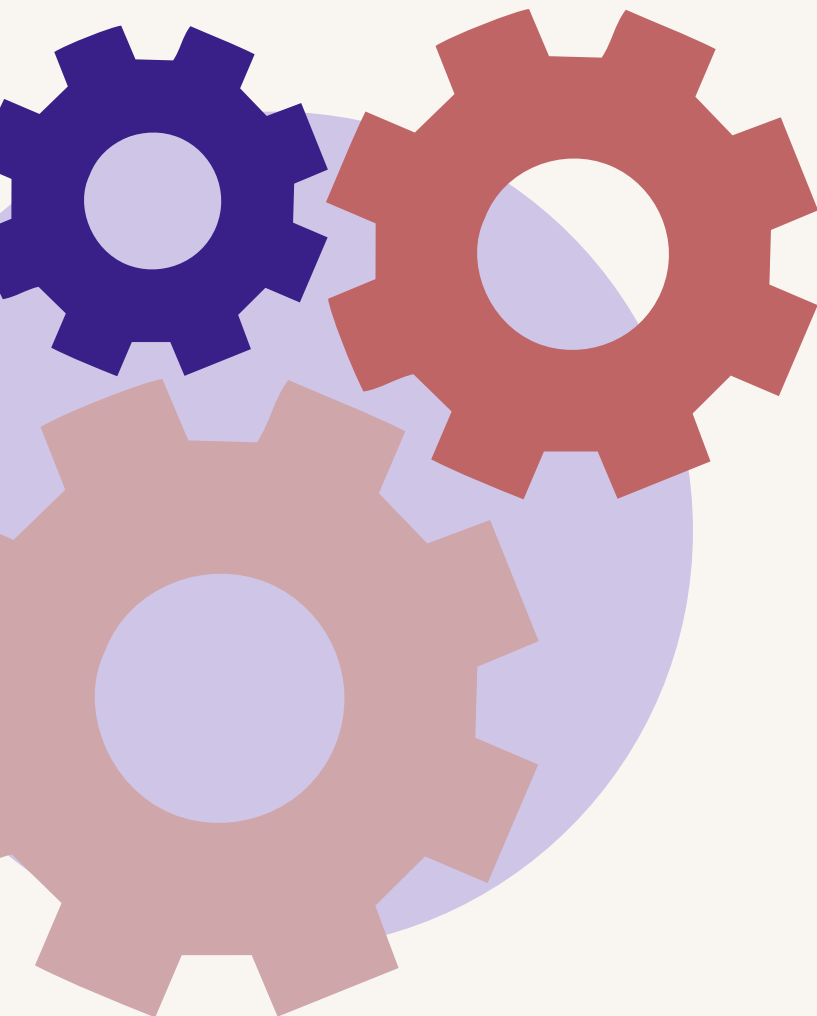
### Policy sub-goals related to transparency:

#### 1 Enable users to interpret the system's output and use it appropriately:

This sub-goal focuses on ensuring transparency for users to interpret AI system outputs accurately. It requires the AI system's operation to be sufficiently transparent (Requirement 1) and the provision of instructions and information for users to operate the AI system correctly (Requirement 2).

#### 2 Prevent risks of impersonation or deception:

The second sub-goal aims to minimize risks associated with individuals unknowingly interacting with AI systems. This includes both intentional deception and situations where individuals are unaware of interacting with AI. To address this, providers must inform users when they are interacting with an AI system (Requirement 3). Additionally, users must inform individuals when exposed to emotion recognition or biometric categorization systems (Requirement 4), and deep fakes must disclose their artificial nature (Requirement 5). This sub-goal will be further discussed in the third part of the EU AIA Open Loop project.





## Theory of Change for risk requirements

The risk requirements in the AIA also contribute to the long-term policy goal of establishing trustworthy AI.



Figure 3. Theory of change for risk requirements in the AIA

### Policy sub-goals related to risk management:

#### 1 AI systems are monitored throughout the entire lifecycle.

This sub-goal focuses on identifying and analysing risks throughout the AI lifecycle. It requires the establishment, implementation and maintenance of a risk management system (Requirement 1). It also requires the identification and analysis of foreseeable risks (Requirement 2). The estimation and evaluation of risks that may emerge when the AI system is in use (Requirement 3), the set up of a post-monitoring system (Requirement 4), the consideration of effects and possible interactions resulting from compliance with Chapter 2 of the AIA (Requirement 5) and the testing made against preliminary defined metrics and probabilistic thresholds (Requirements 6).

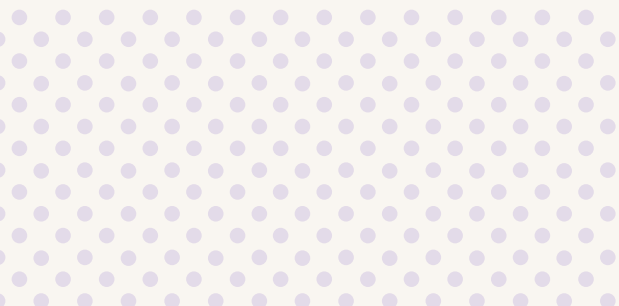
#### 2 Proper risk management measures are in place.

This sub-goal focuses on the adoption of risk management measures throughout the AI lifecycle. The sub-goal is achieved if the following requirements are met: risks are reduced through the design and development (Requirement 7), mitigation and control measures are set up for those risks that cannot be eliminated (Requirement 8), and adequate information and training to users is provided (Requirement 9).

Requirements 4 and 9 are marked with a dashed line, indicating that they apply after the AI system is placed on the market. The remaining requirements should be met prior to placing the AI system on the market.

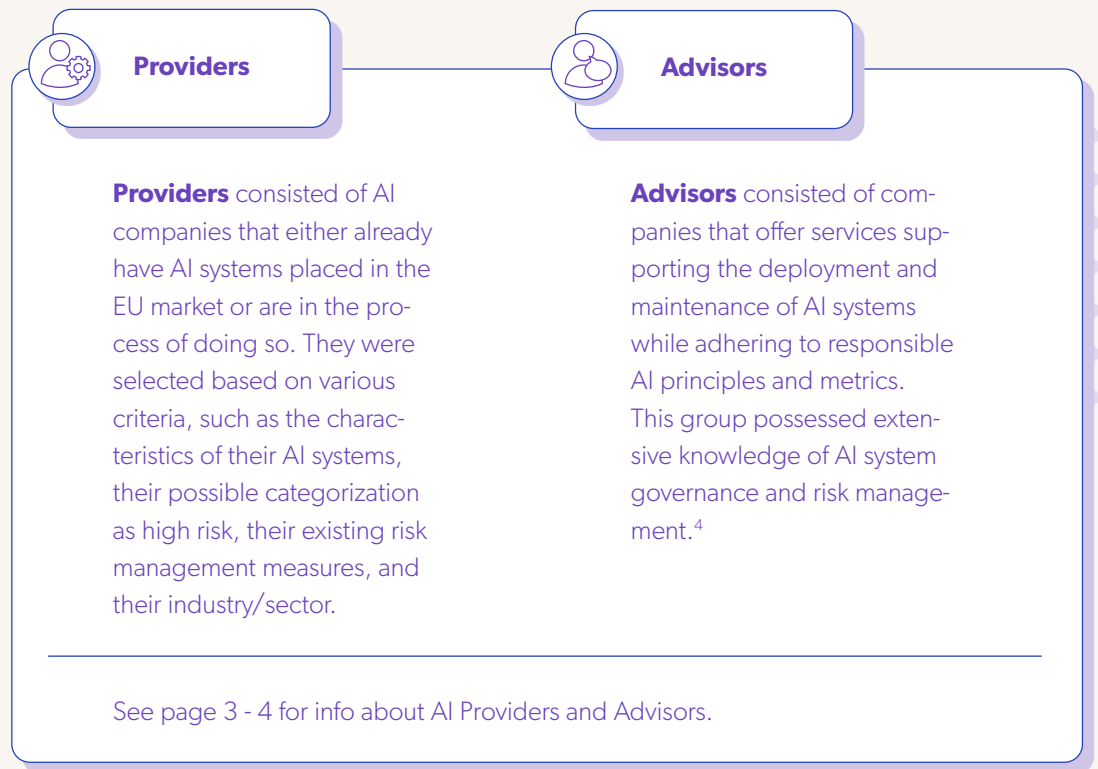
## Program Description: A Policy Prototyping Experiment on the AIA

In this phase of the Open Loop EU AIA program, we selected companies that took part in the Open Loop Forum (OLF)<sup>3</sup> – the first phase of this program – and deep dived into the requirements that companies will have to meet once the AIA enters into force.



## Participants

We engaged two groups of participants in the program: AI providers and AI advisors.



Throughout this report, both providers and advisors are collectively referred to as participants.

## Data Collection

Data collection was conducted through a continuous survey over a period of four weeks. We employed mobile ethnography techniques<sup>5</sup> to gather feedback from participants, a commonly used method in product and service design. The feedback encompassed multiple-choice and Likert scale questions, as well as free-format responses in various formats such as text, audio, video, mind maps, and flowcharts. A detailed explanation of the methodology employed is presented below.

To validate our findings and incorporate diverse perspectives, we organized a co-creation workshop involving experts and stakeholders from the industry, civil society, institutions, and academia. The workshop participants consisted of individuals with expertise in public policy, academia, AI entrepreneurship, and engineering. This workshop focused on three specific topics, which will be described in the following sections.

### 1. Transparency and Human Oversight

Transparency is considered a crucial requirement for establishing trustworthy AI. Specifically, transparency aims to enable users to:

- i) effectively exercise human oversight during the use of AI systems (Article 14), and
- ii) interpret the output of these systems (Article 13).<sup>6</sup>

The EU mandates that providers design and develop AI systems in a way that allows effective oversight by natural persons during system operation. This involves both organizational measures (such as manuals, instructions, and logs) and technical measures (such as explainable AI). Article 14(1) of the AIA states:

*"High-risk AI systems shall be designed and developed in such a way, including with appropriate human-machine interface tools, that they can be effectively overseen by natural persons during the period in which the AI system is in use."*

The workshop focused on enabling effective human oversight, considering the different roles of parties involved (providers and users), determining the target audiences for human oversight measures and interpretability, and identifying strategies to facilitate effective oversight.

### 2. Risk management - Substantial Modification and Pre-determined Changes

The EU is working on the assumption that self-learning AI's will over time change to such an extent that the technical documentation and the conformity assessment are no longer accurate. When an AI system experiences a "substantial modification," the existing conformity test becomes invalid and must be repeated. To avoid overburdening providers, limited changes to the AI system once it is in operation should therefore be allowed.

These potential changes, and particularly their scope, should be determined beforehand. The EU calls this 'pre-determined changes' (see recital 66 AIA). Annex IV para 2 under f requires providers to give:

*“a detailed description of pre-determined changes to the AI system and its performance, together with all the relevant information related to the technical solutions adopted to ensure continuous compliance of the AI system with the relevant requirements set out in Title III, Chapter 2”*

If the change is outside of these parameters, the system is considered to be ‘substantially modified’ and must undergo renewed conformity testing. The ability of providers to define pre-determined changes is a key consideration, as it reflects the feasibility of the EU’s approach to product safety and liability in regulating AI.

The workshop addressed several crucial questions, including strategies for managing the concept of substantial modification in constantly changing AI systems, policy approaches for products that undergo continuous changes, processes for determining the appropriate timing for redoing conformity assessments, and the development of metrics, boundaries, and thresholds for identifying substantial modification.

### 3. Risk Management - Validation and Testing

Providers are obligated to describe the procedures employed to test and validate the accuracy, robustness, cybersecurity, and potential discriminatory impacts of their AI systems, ensuring compliance with the requirements outlined in Title III, Chapter 2 for high-risk AI systems. Additionally, metrics must be defined for measuring accuracy, robustness, cybersecurity, and compliance with the relevant requirements.

Furthermore, metrics for identifying potentially discriminatory impacts, as well as metrics for determining compliance with all the relevant requirements of Title III, Chapter 2 must be developed. Finally, test logs and test reports must be dated and signed by the responsible persons.

During the deep dive activity, we explored participants’ understanding of the metrics used to measure accuracy, robustness, and cybersecurity, as well as their ability to comply with these requirements. **Although all participants recognized the importance of accuracy, robustness, and cybersecurity, they employed different approaches for measuring these factors.**

The workshop aimed to address challenges related to achieving uniformity in measuring and documenting accuracy, robustness, and cybersecurity and explored the possibility of establishing standard metrics.

### Limitations

While conducting the policy prototyping program, several considerations and constraints influenced the scope and depth of the exercise. These factors should be taken into account when interpreting the findings and recommendations presented in this report.

### Participant Representation:

The program involved a limited number of participants,<sup>7</sup> which impacted the ability to achieve a fully representative sample for quantitative analysis.<sup>8</sup> It was a deliberate choice to prioritize qualitative feedback and gather valuable insights from participants, complementing the earlier quantitative inputs. Although this approach has proven effective in generating empirical data and actionable policy recommendations in previous Open Loop programs,<sup>9</sup> it should be noted that the quantitative conclusions drawn from this exercise are limited.

### Time Limitations:

The prototyping exercise was constrained by a relatively short timeline of four weeks. Typically, such exercises require more extensive time frames, ranging from several months to over a year, to facilitate in-depth exploration and analysis. The compressed timeframe limited the level of participant engagement and the opportunity to delve deeply into specific cases. Despite these time constraints, the program team strived to prioritize relevant provisions and focus on key testing goals and assessment objectives.

### Participant Diversity:

The program acknowledges the limitation regarding the diversity of participants. While participants came from different countries, possessed diverse cultural backgrounds, and represented various applications and business models, they predominantly operated as providers of AI systems. To address this potential bias, the program included advisors who provided alternative perspectives as non-providers of AI systems. Furthermore, insights from experts in academia, civil society organizations, and policymakers who participated in the workshop were incorporated to validate and enrich the analysis. However, it is important to recognize that broader diversity among participants could have further enhanced the robustness of the findings.

These considerations and constraints should be taken into account when interpreting the outcomes of the policy prototyping program. Despite these limitations, the program aimed to generate valuable insights and contribute to the ongoing development of effective AI policies. The findings and recommendations presented in this report should be evaluated in light of these factors to gain a comprehensive understanding of the program's outcomes.



# 3

## Part 1: Transparency





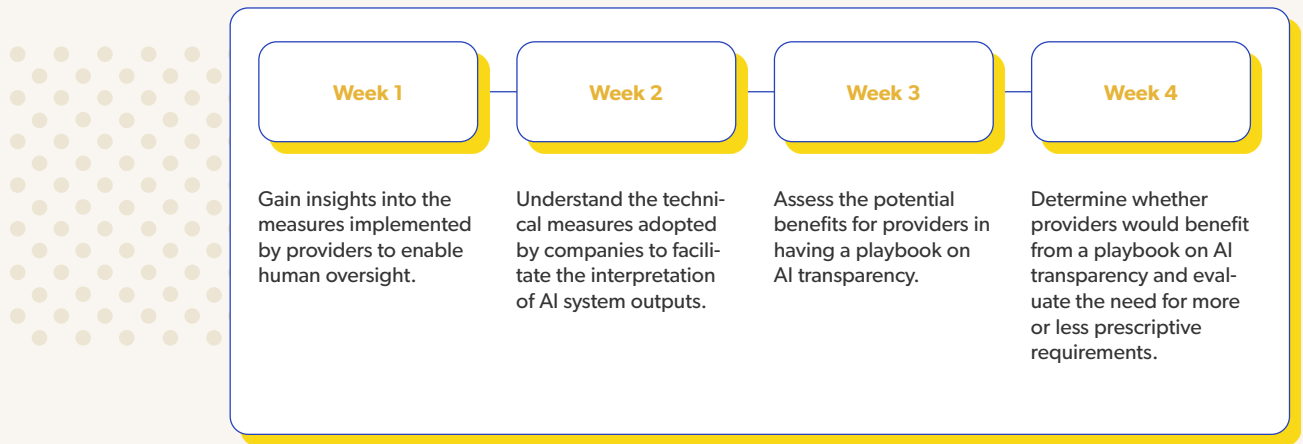
## Introduction

Transparency plays a crucial role in ensuring trustworthy AI systems, as recognized by policymakers. It serves two main purposes:

- i) enabling effective human oversight during the use of AI systems (Article 14), and
- ii) facilitating the interpretation and appropriate utilization of AI system outputs (Article 13).<sup>10</sup>

In addition to these requirements, Article 52 of the AIA emphasizes the need for transparency to ensure that individuals interacting with AI systems are aware that they are engaging with an AI system.<sup>11</sup>

The objective of this deep dive on transparency requirements was as follows:



The purpose of this exercise was to gather valuable insights and inform policymakers on effective strategies for promoting transparency in AI systems.

## Activity 1: Human Oversight (Article 14) in the AIA

Article 14(1) of the AIA mandates providers to design AI systems that can be effectively overseen:

*“High-risk AI systems shall be designed and developed in such a way, including with appropriate human-machine interface tools, that they can be effectively overseen by natural persons during the period in which the AI system is in use.”*

To assess the feasibility and clarity of this requirement, we investigated current human oversight measures and compared them to the AIA's requirements. We also examined the potential impact and cost implications of implementing these requirements.

Our findings provide insights into the organizational measures, the need for guidance or standardization, the division of responsibility, and the challenges related to confidentiality, feasibility, and cost.

## Results

**1 Organizational Measures:** Participants in the program have implemented various organizational measures to enable human oversight. Collaboration with universities, researchers, and subject matter experts is a common practice among participants to ensure diverse perspectives and validation. Documentation and storage of results emerged as a key focus for most participants. They emphasized the following examples of documentation and storage practices:

- i) Documentation of algorithms used.
- ii) Monitoring and storing outcomes generated by their AI systems.
- iii) Exporting result sets<sup>12</sup> for further analysis and discussion.
- iv) Documenting the training process, including information on decisions and features.
- v) Documentation of the development, testing, and outcomes of AI systems.
- vi) Maintaining development roadmaps and tracking their implementation.

In addition, participants highlighted the importance of information sharing and engaging with stakeholders to facilitate discussions around their results. They expressed a commitment to sharing information on specific decisions and features of their AI systems and investing significant time in explaining the systems to clients, ensuring accurate user expectations. Most participants emphasized that these measures primarily target the operators of the AI systems, such as doctors or judges, who use the systems in their daily work. Instructions and guidance provided by participants in the Open Loop Forum were also oriented towards these operators.<sup>13</sup>

Participants demonstrated a proactive approach in implementing organizational measures that promote transparency and facilitate effective human oversight in the use of AI systems.

**2 Human Oversight on Complex Models:** One of the advisors emphasized the significance of effective human oversight, particularly in the case of complex models with a high number of features. Controlling the dynamics and behavior of such models is extremely challenging, making it impractical to rely solely on predefined boundaries or thresholds for future changes. Instead, the advisor suggested that promoting human oversight and interaction is more effective in managing the outcomes of these complex models.

**3 Feasibility and Cost:** The majority of participants considered the cost of implementing human oversight as average. The costs were estimated to vary based on the project's scope and complexity,<sup>14</sup> with estimates ranging from 40 work hours for a small project to 400 work hours for a bigger project.<sup>15</sup> Providers anticipated potential overhead costs, with one participant projecting it to be around 20% of the project budget.



## Observations

Drawing from the aforementioned results, the following observations can be made:

### 1 *Types of Measures:*

There is a wide range of organizational and technical measures mentioned by participants to enable human oversight. However, there is a lack of centralized documentation and standardization of these measures.

This wider array of technical and organizational measures include:

- Technical Measures:
  - Utilizing LIME or SHAP values presented in tabular or graphical formats to explain AI system outputs.
  - Employing standard AI monitor dashboards that display prediction values, historical trends, and sometimes global explainability values.
  - Utilizing explainability methods that are easily understandable by the intended audience.
  - Monitoring input data, model performance, feedback, and other relevant factors.
- Organizational Measures:
  - Conducting assessments of AI/ML systems against responsible AI requirements and generating reports and governance artifacts, including transparency reports, disclosure algorithmic impact assessments, bias audit reports, and risk and compliance reports.
  - Developing dashboards that present pre-identified risks curated by expert panels.
  - Implementing a 10-step manual test or questionnaire to assess risks and assign a risk score based on industry standards.
  - Establishing feedback loops that allow stakeholders to override or approve decisions made by the AI system.
  - Adapting measures to align with the MLOps (Machine Learning Operations) process and ensuring the involvement of the appropriate stakeholders.

These observations highlight the breadth of measures implemented by participants, encompassing both organizational and technical approaches, to facilitate effective human oversight in AI systems.

Given the broad nature of organizational (and technical) measures available and the absence of efforts in centralizing, documenting and making available these measures, it might be beneficial to create some form of guidance or standardization as to which human oversight measures are needed, taking into account the risk posed by the AI system. Answers by the participants and advisors have shown that **guidance or standardization should focus both on technical and organizational measures to enable human oversight. Organizational and technical measures are complementary and are mutually essential to enable human oversight.** For example, (research) scientists or subject matter experts are essential for the technical measures to fit the context.

## 2 *Division of Responsibility for Human Oversight:*

The division of responsibility for human oversight is not clearly defined in the AIA. While the provider is mandated to facilitate human oversight, it is unclear who is responsible for overseeing the AI system in real-world applications.

Participants believed that both the provider and user should share the responsibility, emphasizing the need for a clear delineation of responsibilities.

6 out of 7 participants in our exercise believe that both the user and the provider share the responsibility of human oversight. Our participants, who are AI system providers, have their own internal human oversight measures as part of their MLOps process, and also provide measures to help users interpret AI system results.<sup>16</sup> In the first part of our AIA policy prototyping experiment, we found that participants often consider themselves as both providers and users because they use AI components to operate their own AI systems or incorporate other AI systems into their own.<sup>17</sup> In both situations, whether a user is also a provider or solely a user, it is beneficial to have a clear delineation of responsibilities.

To ensure effective human oversight, it is advisable to provide further clarification on rules related to human oversight, particularly regarding the allocation of responsibility for this task in the AIA or associated subordinate regulation/guidance.

### 3 **Confidentiality and Accessibility:**

Confidentiality can pose a potential obstacle to third-party human oversight, whether it is carried out by users or regulators. During the deep dive sessions, a participant noted that they were unable to provide extensive details about their organizational measures due to confidentiality concerns. While this limitation is understandable within the context of the exercise, it also applies to real-life situations where users may face similar constraints.

Certain organizational measures related to human oversight, such as identifying the specific author of a code segment or its implementation location, may be less suitable for sharing with third parties, including users.

### 4 **Feasibility and Cost:**

Feasibility and cost are project-dependent, but most participants considered the requirement feasible with an “average” burden on organizations. However, it is important to note that implementing human oversight measures may result in significant overhead costs for providers (with one participant projecting the cost to be 20% of the budget of a given project to design human oversight measures).

In conclusion, to ensure effective human oversight in AI systems, it is recommended to establish guidance or standards for human oversight measures, clarify the division of responsibility between providers and users, address challenges related to confidentiality and accessibility, and consider the feasibility and cost implications of implementing these measures. These steps will contribute to fostering trust, accountability, and transparency in AI systems, aligning with the goals set forth in Article 14 of the AIA.

## **Activity 2: Transparency (Article 13)**

High-risk AI systems must adhere to the guidelines set forth in Article 13(1) of the AIA, ensuring their design and development promote transparency for users to effectively interpret and utilize the system's output.

Transparency is further emphasized in Article 13(2), mandating that accompanying instructions be relevant, accessible, and comprehensible to users of the AI system. These instructions include a description of the technical measures implemented to facilitate user interpretation of AI system outputs, as stipulated in Article 13(3)(d) and Annex IV(4)(c) of the AIA.

To evaluate the feasibility and clarity of this requirement, we asked participants the type of models they use, the interpretation tools they have put in place to enable transparency, whether they have put in place instructions for the user, their confidence in the effectiveness of these measures to enable users to have a sufficient level of oversight; the burden (i.e., cost) on the organization to provide these instructions and tools.

## Results

**1 Different approaches to enhance interpretability of AI models:** Participants reported employing a combination of black box models and intrinsically interpretable models.<sup>18</sup> For the intrinsically interpretable models, various methods were utilized to provide users with oversight over the system. These methods included:

- The use of Shapley Additive Explanations (SHAP),<sup>19</sup>
- Enabling human oversight by explaining how scores were generated and providing biased examples, and
- offering a general overview of the earning model along with its status and metrics.

Regarding black box models, participants acting as providers are still exploring the most effective approaches to enhance interpretability. However, participants acting as advisors suggested methods such as utilizing model cards,<sup>20</sup> conducting impact assessments and transparency reports, employing techniques like LIME or SHAP in conjunction with prediction scores, and interpreting simpler models to derive insights into black box models.

**2 Target for interpretability methods varies based on context and MLOps process:** Both participants acting as providers and advisors highlighted that the target audience for interpretability methods varies depending on the specific context and the organization's MLOps (Machine Learning Operations) process. The intended audience may include solution engineers, project managers, AI portfolio/risk managers, or operators. The same goes for our group of advisors. It depends on the customer, the use case and the MLOps process. One of the advisors noted: *"Many different people, depending on the MLOps process. There is not a default target group"*.

**3 Customized instructions and diverse methods used by providers to educate users/operators of AI systems:** All participants indicated that they provide instructions to users or operators on how to use their AI systems. The nature of these instructions varies, tailored to the specific AI systems being deployed. Participants mentioned several examples of instructing users, including manuals, guides, how-to's, and FAQs. These instructional materials cover a range of topics, such as screen flows, exporting predictions, accessing methods/data, verification tools, using dashboards, shadowing (observing users and providing feedback), one-on-one instructions, and user training.

**4 Confidence:** The majority of participants expressed confidence in the organizational and technical measures implemented to enable users to exercise sufficient human oversight. Four out of six participants reported being confident, while one participant expressed very high confidence in their measures.

- 5 **Costs and Burdens:** The costs associated with meeting this requirement vary depending on the size and complexity of the project. 3 out of 5 participants rate the burden of providing instructions and interpretability of output as high, one participant as average, and one as low. Estimates ranged from 80 hours total to 300 hours per month for providing instructions and interpreting AI system results. Additionally, participants acknowledged the need to hire external expertise, such as lawyers or consultants, to support the implementation of interpretability measures.

## Observations

Drawing from the aforementioned results, the following observations can be made:

### 1 *Different audiences for human oversight and interpretability:*

The participants' responses indicate that there are distinct audiences for human oversight and interpretability of AI system outputs. Some measures focus on enabling oversight by the provider's internal teams, while others target the operators/users of the AI systems. It is unclear to what extent there needs to be continued involvement of the provider when it comes to monitoring the use of an AI system. Unlike traditional products regulated by product safety regulations, AI systems require ongoing interaction between providers and users during their operation. Users may face challenges in detecting model drift without the support and involvement of the AI system provider.

### 2 *Different methods for interpretability:*

The diversity of the participants' responses highlights the existence of various approaches to providing instructions and interpretability methods. This diversity stems from the fact that AI systems often require specific instructions tailored to individual projects. The EU AIA allows for this flexibility by not imposing strict methods. At the same time though this reduces legal certainty as it is unclear, yet which methods are sufficient (a common issue with principle-based legislation).

### **3** *The maturity of AI companies appears to influence the choice of methods:*

In the early stages, when AI companies are still exploring and developing their systems, providers are closely involved with users, allowing for direct communication and customized instructions. However, as AI companies grow and introduce more products to the market, maintaining such close involvement becomes challenging. Instructions need to be more generalizable and less reliant on constant provider-user interaction.

### **4** *A potential approach worth exploring is a modular approach to instructions,*

which combines a hands-on approach when feasible with standardized guidelines. This would preserve the diversity of interpretability methods while providing guidance to providers and users on the recommended level of interaction. Such methods could facilitate both customization and interoperability.

In conclusion, our findings demonstrate the range of approaches adopted by participants to enhance transparency and interpretability in AI systems. While there are different audiences for oversight and interpretability, providers strive to offer customized instructions and educate users/operators through various means, including manuals, guides, FAQs, shadowing, and training. Participants expressed confidence in their measures but highlighted the high burden, both in terms of costs and external expertise required. To ensure legal certainty, it may be beneficial to explore a modular approach that combines hands-on interaction and standardized guidelines, promoting interpretability while accommodating diverse AI system contexts.

## **Activities 3 & 4: The benefits of technical guidance**

During the Open Loop Forum on the AIA, it was concluded that providers would benefit from guidance on complying with the AIA requirements. To explore this further, we provided participants with a guidance on AI transparency<sup>21</sup> and assessed its impact on their AI systems' transparency and interpretability.

## **Results**

- 1** **Technical Guidance increases focus on interpretability and explainability.** Participants overwhelmingly expressed that they would implement additional measures after reading the technical guidance, particularly in how they present the use and results of their AI systems to users. These measures aimed to enhance the interpretability and explainability of the AI



systems. The guidance also inspired participants to develop new methods and workflows to improve transparency, such as using templated workflows and user surveys to assess the effectiveness of explanations.

- 2 Technical Guidance promotes awareness and action on explainability methods in AI systems.** The technical guidance raised awareness about the presentation of explainability methods to different actors and emphasized the need for tailored approaches. Participants expressed satisfaction with their current explainability methods but acknowledged the importance of improving how they present AI system information to users.<sup>22</sup> They recognized the value of evaluation approaches and expressed the desire to implement user surveys to gauge user understanding and the impact of provided information on their actions and decisions. The *Newsroom* mentioned: “We need better evaluation approaches for transparency, to assess the effectiveness of the explanations and their effect on mitigating risks.”
- 3 Participants recognize the need to map out the needs of different actors.** Participants became more aware of the importance of explainability for different actors and suggested mapping the needs and expectations of these actors, including end users. They acknowledged that explainability methods vary depending on the fields and algorithms used, making general guidance complex. An example provided was the need to account for the potential impact and significance of an AI system used in medical decision-making. Participants emphasized the importance of collaboration with UX/UI designers to strike the right balance between providing information without overwhelming end users.
- 4 Increased emphasis on the end user as a target audience.** Participants realized the need to place more emphasis on the end user as an audience target. They acknowledged that while key algorithms may be explainable to AI managers, they should also be transparent to the end user. Participants suggested providing more information to the end user on how the AI system makes decisions and expressed interest in evaluating the effectiveness of transparency methods through surveys.
- 5 Cost and burden implications.** Although participants had difficulty quantifying the costs of transparency and explainability measures, they estimated that implementing additional measures would have a medium to high cost impact.
- 6 Technical guidance facilitates compliance with AIA requirements.** All participants agreed that the technical guidance provided would help companies comply with the AIA requirements, particularly those outlined in Articles 13 and 14. They highlighted that technical guidance saves time and effort by offering clear instructions, best practices, and a starting point to translate requirements into specific actions. It can also help smaller organizations with limited resources quickly understand and apply what is required.

Participants listed several reasons for why a technical guidance would be beneficial:

### Telesoftas:

*"(Small) organizations have limited time and resources to conduct research or develop measures. Technical guidance can help them quickly understand what is required and how it can or should be applied."*

### Synerscope:

*"Technical guidance can help alert organizations on the necessity of these measures. The technical guidance made the participants realize and better understand the importance of these measures."*

*"Technical guidance can help to be compliant with specific requirements of the AIA."*

### The Newsroom:

*"Legislation as the AIA is often long and unstructured. A technical guidance with a step-by-step guide brings structure and makes it easier to translate requirements to specific actions."*

### DLabs:

*"Technical guidance can provide a starting point and a set of best practices. This avoids AI companies from re-inventing the wheel. The technical guidance can save AI companies a lot of time and effort by providing a map on where to start and how to approach this legislation."*

## Observations

Based on the aforementioned results, the following observations can be derived:

1

The main observation from this activity is that participants found the technical guidance beneficial. They noted that the AIA legislation can be difficult to interpret and lacks clarity in some parts. In particular, the 'nested requirements' of the AIA may prove to be difficult to understand and implement. The AIA has a structure whereby many of the requirements (e.g. those in chapter 2) are further fleshed out in the technical annexes. These annexes include a high level of detail and oftentimes refer back to the Articles in the AIA itself, making reading and understanding them more difficult. One of the participants argued:

*"The AIA is quite long and unstructured. The technical guidance provided is very well structured and provides a useful step-by-step guide. This makes it easier to translate requirements to specific actions we must take as a company to comply in a way that is also constructive and positive for our business."*

- The Newsroom

The structured approach of the guidance helped participants better understand the requirements and suggested that official guidance alongside the AIA could enhance its effectiveness. Participants also appreciated the concrete and practical nature of the guidance, which raised awareness about the ethical and legal implications of AI more effectively than abstract requirements in the law.

In order to improve the effectiveness of the AIA the legislator could contemplate issuing official guidance alongside the AIA.

2

Another observation highlighted by a participant is the varying level of specificity in the AIA's requirements. While certain aspects are highly prescriptive, others remain vague and open-ended.<sup>23</sup> For instance, whereas the requirement of Annex IV (2) under (d) is very prescriptive:

*"where relevant, the data requirements in terms of datasheets describing the training methodologies and techniques and the training data sets used, including information about the provenance of those data sets, their scope and main characteristics; how the data was obtained and selected; labelling procedures (e.g. for supervised learning), data cleaning methodologies (e.g. outliers detection);*

requirement of Annex IV(2) under (f) is quite vague and open-ended:

*"where applicable, a detailed description of pre-determined changes to the AI system and its performance, together with all the relevant information related to the technical solutions adopted to ensure continuous compliance of the AI system with the relevant requirements set out in Title III, Chapter 2;"*

This lack of consistency can be addressed by complementary guidance that provides additional clarity and interpretation. Striking the right balance between legal certainty and flexibility is challenging, and guidance can help navigate this complexity.

**3**

Furthermore, the technical guidance raised awareness among participants regarding the ethical and legal implications of AI. The concrete approach and specific language used in the guidance resonated more effectively with providers and users compared to the abstract requirements outlined in the AIA. For instance, the guidance offered explicit examples of explainability methods, whereas the AIA emphasizes the need for instructions to be "concise, complete, correct and clear information that is relevant, accessible and comprehensible to users" (AIA Article 13).

In conclusion, the observations underscore the significance of technical guidance in facilitating compliance, addressing interpretational challenges, providing clarity, and raising awareness about ethical considerations in the context of AI policy. Integrating complementary guidance alongside the AIA can further enhance its effectiveness and enable stakeholders to navigate the complexities of AI regulation more effectively.

# 4

## Part 2: Risk Management in the AIA



## Introduction

The AIA aims to address increasing adoption and use of AI-systems in the EU. The European Commission recognizes the added value of such systems but also identifies risks related to this development. The risk management requirements in the AIA are aimed at reducing the risks of AI systems. Article 9 of the AIA introduces the requirement for providers of high-risk AI systems to have a comprehensive risk management system in place throughout the lifecycle of the AI systems.

The risk management system must, among other things: (i) identify known and foreseeable risks associated with the AI system; (ii) estimate and evaluate risks that “may emerge” when the system is “used in accordance with its intended purpose and under conditions of reasonably foreseeable misuse”; and (iii) include risk management measures. Residual risks, the risks that are not directly covered by the mentioned measures, must be of an “acceptable” degree and communicated to users. Among other requirements, high-risk AI systems must also be tested prior to being placed on the market or put into service to identify the most appropriate risk management measures.

### Activity 1: Describing the development of the AI system

The AIA mandates that AI providers describe the methods and steps taken in the development of their AI systems (see Annex IV para 2 under a). This requirement serves the purpose of enabling regulators and notified bodies to assess compliance. Annex IV, paragraph 2(a) requires AI providers to describe: “the methods and steps performed for the development of the AI system, including, where relevant, recourse to pre-trained systems or tools provided by third parties and how these have been used, integrated or modified by the provider”.

The purpose of this exercise was to evaluate how users describe their AI systems and assess whether these descriptions are adequate for regulators and notified bodies to determine compliance. Additionally, the assessment aimed to gauge the administrative costs associated with compliance.

To gather this information, participants were asked a series of questions, including their ability to describe and document the development process, their use of third-party systems or tools,<sup>24</sup> the sufficiency of the provided information for assessing compliance, the ability to describe the development process without revealing sensitive information,<sup>25</sup> and their assessment of the task of identifying and documenting the required information.

## Results

**1 Methods and steps used in developing an AI system.** All participants were able to describe the methods and steps used in developing their AI systems, following a comparable process. The process can be described as follows:

- i) Define use case, functionality of the system, or problem identification.
- ii) Assess availability of data, basic data exploration and preparation of initial data set.
- iii) Design AI models, and define methodology for validating the AI or ML model.
- iv) Build the product, integrating AI models into the overall product and system, and training ML-based AI models. Evaluate the performance.

- v) Design documentation and processes required for use and compliance.
  - vi) Deploy, test, and iterate.
  - vii) When it comes to documenting all the activities within these steps in detail, participants noted that this is a hard task.
- 2 **Pre-trained models and third-party tools.** Several participants integrated pre-trained systems or tools from third parties into their AI systems. One participant, for example, utilized a large language model (BERT<sup>26</sup>) to predict the association between skills and learning content.
  - 3 **Relevance of the information.** The goal of this requirement is to enable regulators to assess how the AI systems are developed. Although the participants considered the requirement somewhat sufficient and relevant, they highlighted that the regulator should require explicit information on the context in which the AI system is used, to better understand its development. A participant acknowledged the importance of analyzing the method and steps for the development of AI, for instance by inspecting the data or pre-trained models that were used to build a new model. However, the only reliable way to assess a model’s performance is to use and monitor it on a new data set or by getting access to past performance data.
  - 4 **Trade secrets and sensitive information.** Out of the five participants, four could not describe how their AI systems are developed without divulging trade secrets or any other sensitive information.<sup>27</sup>
  - 5 **Administrative burden.** Fully identifying, documenting, and maintaining the required information posed a significant administrative and compliance burden for participants. The estimated impact on organizations in terms of costs and burden was substantial, but difficult to quantify due to the ever-changing nature of AI systems and evolving regulatory requirements. Participants suggested the need for efficient and automated documentation processes that combine qualitative and metric information, but acknowledged the challenges in labeling and incorporating GDPR requirements.

## Observations

Based on the aforementioned results, the following observations can be derived:

### 1 **Administrative Burden and Relevance for Regulators.**

Participants possess the capability to describe and document the steps involved in developing AI systems. Nevertheless, this documentation process places a notable burden on them. While the information provided is deemed somewhat sufficient and relevant by them, capturing the necessary contextual information within technical documentation remains a challenge.

## **2** *Importance of Efficient and Automated Documentation.*

Efficiency and automation are crucial in achieving comprehensive documentation. Panelists emphasize the need for an approach that integrates qualitative information with metric information, combining automated documentation with human review. However, manual labeling, which is not commonly practiced in governments, poses affordability challenges. Additionally, the inclusion of GDPR requirements further complicates the documentation process.

## **3** *Concerns about Compliance Costs and Effectiveness.*

The perceived high cost of compliance and potential low relevance in assessing compliance raise concerns about the effectiveness of documentation requirements. Evaluating compliance solely based on documentation can be burdensome and may not provide accurate insights into AI system performance.<sup>28</sup>

## **4** *Addressing Third Party Models and Tools.*

A significant point of attention lies in the use of third-party models, tools, and data in non-high risk applications. Most participants heavily rely on these external components. Hence, it is crucial to determine the extent to which technical documentation requirements encompass third-party contributions. Participants face challenges in complying with AI regulations due to limited access to technical details and documentation of third-party models. Two potential solutions are proposed: exempting third-party models from documentation requirements, risking the omission of important information, or mandating third-party providers to issue technical documentation to supervisory authorities and clients (namely AI users).

## **5** *Protecting Trade Secrets and Sensitive Information.*

Participants raise concerns about documenting their development process without revealing trade secrets or sensitive information. While this issue is not unique to AI regulation, it must be carefully considered when handling technical documentation. Supervisory authorities must ensure that public versions of findings and decisions do not disclose trade secrets or sensitive details.



In sum, documentation requirements pose significant challenges in the development and compliance of AI systems. Balancing administrative burdens, automation needs, consideration of third-party contributions, and protection of trade secrets is vital for effective regulatory oversight. Policymakers should address these concerns to foster innovation while ensuring transparency and accountability in AI development and deployment.


## Activity 2: Pre-determined changes

The EU is working on the assumption that self-learning AI's will over time change to such an extent that the technical documentation and the conformity assessment are no longer accurate. When an AI system undergoes substantial modification, the conformity test becomes invalid and must be repeated. To avoid burdening providers, limited changes to the AI system should be allowed once it is operational. These allowable changes, known as pre-determined changes, need to be defined in advance and documented. If changes go beyond these predetermined parameters, the system is considered substantially modified and requires renewed conformity testing.

Providers' ability to define pre-determined changes is crucial for assessing the feasibility of the EU's product safety and liability approach to regulating AI. This approach treats AI systems as off-the-shelf products with assessable predetermined changes and a taxonomy based on product and safety liability.

To test this requirement, participants were asked about substantial modifications, defining pre-determined changes, the feasibility of determining changes, and the impact on their organizations in terms of cost and burden.

## Results

- 
- 1 **Uncertainty about the term 'substantial':** Participants lacked a clear understanding of what constitutes a substantial modification. While some considered important enhancements to be substantial modifications, they argued that the core purpose of the system would remain unchanged from the user's perspective.
  - 2 **Defining pre-determined changes:** Participants provided various examples of pre-determined changes, mainly aimed at improving the AI system's functioning rather than introducing new functionality or altering existing functionality. *Learnershape* provided the following examples of pre-determined changes: “(1) *improving relevance of recommended content*, (2) *recognizing the level (i.e. sophistication) of content* and (3) *recognizing the quality of content*.” *Peregrine AI* gave the following examples: “*Higher accuracy of object detection, new object classes, new parameters the system can use to interpret visual data in real time*.”
  - 3 **Feasibility of pre-determining changes:** While participants believed it was possible to define pre-determined changes, they stated that they often change or correct their way of working based on new insights and learnings during the development process. As such, it might hurt the agility of the companies to follow pre-determined changes, due to the lack of possibility to adjust course, based on new learnings. They also highlighted the difficulty of predicting technological progress and changes in use cases, which could hinder the agility of companies in adapting their AI systems. Next to the participants, the advisors also agree that it is nearly impossible to cover all potential changes. The advisors identified the possibilities

to define boundaries/thresholds for systems that are relatively static and do not undergo continuous change. However, many AI systems are frequently or constantly retrained, creating an extremely dynamic system, which makes it exceptionally hard to define their boundaries/threshold. Therefore, it could be easier to set a minimum adequacy bar for aspects like performance.

- 4 Cost and burden:** Participants rated the impact of defining and monitoring pre-determined changes as average to high, but quantifying the costs at this point was challenging for them.

## Observations

Based on these results, we can elaborate the following observations.

- 1 *The requirement to define pre-determined changes is pivotal but somewhat hidden in the annexes of the AIA.***

Failure to pre-determine changes could result in any modifications being considered substantial, leading to significant compliance burdens as AI systems evolve over time.

- 2 *The terms 'substantial modifications' and 'pre-determined changes' lack clarity in the context of AI systems.***

Different interpretations exist, with some viewing all improvements as pre-determined changes and others associating it only with new functionality. Recording reasonable expectations regarding system changes may help increase provider awareness and allow for objective assessment.

- 3 *Participants' agile development approach, adjusting course based on new insights and learnings, complicates the determination of system changes in advance.***

AI systems and the development process are not static, posing challenges to defining pre-determined changes. The concept of pre-determined changes raises concerns given the AIA's focus on mandatory conformity assessment. Workshop findings emphasized the need for general principles and standards applicable to all AI systems at both the system and model levels.

## 4

***The rapid and continuous improvement of AI systems is essential to their success.***

However, these changes may trigger the requirement for conformity assessment when reaching substantial modification. Clarification from the legislator on what constitutes substantial modification in the AI context and how to measure and define it, such as through boundaries or thresholds, is necessary.

In conclusion, the ambiguity surrounding terms such as "substantial modifications" and "pre-determined changes" raises concerns regarding compliance burdens and varying interpretations among users. The agile nature of AI development further complicates the ability to determine in advance how systems will change. The concept of pre-determined changes proves problematic, considering the mandatory conformity assessment framework of the AIA. The need for general principles, standards applicable to all AI systems, and clearer guidance from the legislator regarding substantial modifications becomes evident. Furthermore, further clarification is required to define what constitutes a substantial modification and how it should be measured and assessed within the context of AI. This clarification is essential to strike the right balance between enabling the rapid improvement and adaptability of AI systems while ensuring effective regulatory oversight.

**Activity 3: Monitoring**

Post-market monitoring is a crucial requirement of the AIA to address emerging risks from AI systems that continue learning after deployment (Article 61 AIA). While not explicitly mentioned, monitoring the technical performance of AI systems is an integral aspect of post-market monitoring. Providers are mandated to implement logging capabilities for monitoring high-risk AI systems (Article 12(3) AIA). In this activity, we asked participants whether they monitor the correct operation of their AI systems after deployment and, if not, why. We investigated which elements they monitor and we asked who is responsible to monitor the performance of the model.



## Results

- 1 Monitoring.** Most participants monitor the correct operation of their deployed AI systems. One participant cited limited access to the customer's tenant as the reason for not monitoring, as decided by the client for security reasons. Regarding monitored elements, participants focus on various aspects. All participants monitor the quality of input data used for training. Additionally, most participants monitor data and feature drift<sup>29</sup>, model drift, and model configuration. Elements monitored for output include model performance in production<sup>30</sup>, model input/output distribution<sup>31</sup>, model training and re-training<sup>32</sup>, model evaluation and testing<sup>33</sup>, hardware metrics<sup>34</sup>, CI/CD pipelines for ML<sup>35</sup>, accuracy of predictions and classification<sup>36</sup>, and data mismatch, integrity, and drift<sup>37</sup>. One advisor emphasized the effectiveness of monitoring input data rather than the dynamics of the model itself or the output. They recommended continuous explanation of the model's prediction process and incorporating user feedback to steer models in the right direction:

*“Continuously explain as best as possible how a model comes to a prediction, and provide methods for feedback by end users who can in this way steer models in the right direction. Often adverse outcomes arise when models seem to run within boundaries, but developers have simply overseen some of the design consequences.”*

– Deploy

- 2 Responsibility for monitoring.** Four out of five participants consider it their responsibility as providers to monitor the performance of their models. One participant believes it is a shared responsibility between the provider and the user due to their collaborative work on the AI solution. Participants see it as their responsibility because they are the providers and need to ensure high-quality and efficient AI solutions. The client's limited access to the AI system further reinforces the providers' responsibility. The participant advocating shared responsibility highlighted the interaction and cooperation between providers and users.
- 3 Costs and burden.** Regarding the burden of monitoring, all participants perceive it as average. Estimating costs proved challenging, with one participant approximating it between 10% and 20% of the total data science effort.
- 4 Monitoring thresholds and boundaries.** An advisor described monitoring predetermined boundaries by assessing predictive quality and explainability. They set lower bar thresholds based on previous training models and predictive parameters like Accuracy Score or F1 score. Additionally, they analyze the order and explainability scores of the top 5-10 features, detecting changes and variations.

## Observations

**The current division of responsibility in the AIA lacks clarity and could benefit from further clarification.** Responsibility for monitoring the AI system operation depends on how the system is placed on the market. When offered as an end-to-end solution with promised outcomes, participants suggest making the provider responsible. However, in cases involving close interaction between providers and users, shared responsibility or shifting responsibility to the user may be more appropriate.

## Activity 4: Validation and testing

Providers are obligated to describe the methodologies employed to assess the accuracy, robustness, cybersecurity, and potential discriminatory impact of their AI systems, while complying with the relevant requirements outlined in Title III, Chapter 2.

In this activity, the participants were asked about their practices regarding measuring accuracy, robustness, and cybersecurity of their AI systems. Additionally, the existence of industry standards or best practices for measuring these factors was explored, along with how providers gather information on these metrics. The participants were also asked about the challenges in identifying and documenting metrics for accuracy, robustness, cybersecurity, and discriminatory impacts. Lastly, their overall thoughts on the level of prescription in enabling human oversight according to the EU were requested.

### Results

1 **Accuracy.** Most participants measure accuracy using various approaches, such as:

- Classification accuracy<sup>38</sup>
- Logarithmic loss<sup>39</sup>
- Confusion matrix<sup>40</sup>
- Area under curve<sup>41</sup>
- F1 score<sup>42</sup>
- Mean absolute error and mean squared error<sup>43</sup>
- Measuring AUC of model predictions against human-labelled test sets<sup>44</sup>
- Using related metrics that take different perspectives on the data and outcomes<sup>45</sup>

The identification and documentation of accuracy metrics were perceived as moderately complex, as the choice of metrics depends on the specific AI system and its intended purpose. Some participants faced challenges in defining accuracy metrics for healthcare systems due to the involvement of healthcare professionals and the need to establish thresholds for early detection of cognitive impairment.

2 **Robustness.** Most participants measure the robustness of their AI systems by e.g. generating or modifying inputs and monitoring testing coverage to detect anomalies. They document any defects or unexpected behavior exhibited by the model. Measuring robustness was considered moderately complex, with participants acknowledging the absence of consensus on standard metrics.

One of the advisors (Armilla AI) suggested metrics such as model weakness<sup>46</sup>, decision boundary detection<sup>47</sup>, scenario testing<sup>48</sup>, perturbation testing<sup>49</sup>, and behavioral testing<sup>50</sup>.

Other metrics suggested by another advisor (Enzai) included feature space partitioning<sup>51</sup>, ensemble modeling based on partitioning<sup>52</sup>, distribution shift<sup>53</sup>, and random dropout for neural networks<sup>54</sup>.

- 3 **Cybersecurity.** The majority of participants measure the cybersecurity of their AI systems. Anomaly detection was highlighted as a baseline approach, and some participants assessed the cybersecurity of the overall solution, including the AI system, UI/UX, and other elements. Logging and reviewing system events were mentioned as the primary method, alongside the adoption of best practices to minimize the attack surface. Few participants were aware of industry standards or best practices for measuring cybersecurity, with some relying on general methodologies like CRISP-DM<sup>55</sup>.
- 4 **Discriminatory impacts.** Participants expressed varying opinions on the complexity of identifying and documenting metrics for discriminatory impacts. Most of the participants agree on the fact that data is inherently biased and overcoming this bias is a hard task: *"(...) the identification of biases metrics will be hard as we not only need to make sure that our input data won't introduce biases when training the model, but it will also require ensuring that potential already existing medical and cultural biases are not affecting the training data set of our model."* (Virtuleap) They suggested a human-in-the-loop approach as a potential solution.

## Observations

The following observations can be derived from the highlights above.

### 1 **The importance of accuracy, robustness, cybersecurity, and mitigating discriminatory impacts in AI systems was evident among the participants.**

However, the metrics used for measuring these aspects varied due to the specificity and context-dependence of each system.

### 2 **The lack of clearly defined standards and methodologies for robustness and cybersecurity was noted.**

Additionally, participants emphasized the need to determine where to measure these aspects effectively, as measuring overall robustness may not always yield meaningful insights. To promote uniformity, clear metrics and standards should be developed.

### 3 **The workshop highlighted the importance of considering human rights and dignity in AI and the necessity of reaching a consensus on risk and success metrics for validation and testing.**

**5**

# **Conclusion and recommendations**



Through the exercise of policy prototyping and analysis, this study has identified several key elements of the AIA's Risk Management, Transparency and Human Oversight requirements that require further review to achieve clarity and feasibility. The findings highlight areas where additional guidance and clarification are necessary to ensure the practical usability of the regulation in real-world scenarios.

## Conclusions

### 1 *Human Oversight*

Efforts must be made to centralize, document, and make available the organizational and technical measures for effective human oversight. Clarification is needed regarding the division of responsibility for human oversight.

### 2 *Transparency*

Different audiences and interpretability methods require a modular approach to providing instructions for AI system outputs while allowing for standardization.

### 3 *Technical Guidance*

Concrete technical guidance should be provided to address the complexity of the AIA, which may resonate more with subjects of the regulation.

### 4 *Documentation Burden*

Balancing the need for describing and documenting AI system development without divulging trade secrets or sensitive information is challenging. Third-party models could be exempted from the requirement, or technical documentation should be provided by third-party providers for compliance purposes.



## **5** *Substantial Modifications*

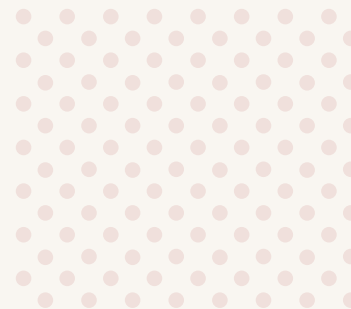
Clarity is needed from legislators on what constitutes a substantial modification in the context of AI systems, along with clear metrics and definitions for measurement.

## **6** *Monitoring Responsibility*

Further clarity is needed regarding the division of responsibility for monitoring AI systems based on how they are placed on the market.

## **7** *Accuracy, Robustness, Cybersecurity, and Discriminatory Impacts*

Clear metrics and standards are necessary to ensure uniformity in measuring these factors during validation and testing.



## Recommendations

Based on the study's conclusions, the following recommendations are proposed to address the identified challenges and improve the effectiveness of the AIA:

### **1** *On Transparency (Article 13)*

The study suggests exploring a modular approach to providing instructions for AI system outputs. This approach would enable a hands-on approach while allowing for standardization, catering to different audiences and interpretability methods.

### **2** *On documenting the development of the AI system*

The study recommends considering two options. Firstly, the requirement of describing and documenting could be waived for third-party models to alleviate the burden on providers. Alternatively, third-party providers should be mandated to issue technical documentation on their models to supervisory authorities and clients for compliance purposes, striking a balance between transparency and protecting trade secrets.

### **3** *On substantial modifications and predetermined changes*

The study highlights the need for further clarification from legislators on what constitutes a substantial modification in the context of AI systems. Clear metrics and definitions should be established to guide measurement and decision-making in this area.

In conclusion, this study's findings shed light on areas within the AIA that require attention and refinement to enhance the regulation's practical application. By implementing the above recommendations, policymakers can address the identified challenges and ensure the AIA is better suited for addressing the complexities of AI systems in real-world settings.



# Endnotes



- 1 Andrade, Norberto Nuno Gomes de, and Antonella Zarra. "Artificial Intelligence Act: A Policy Prototyping Experiment: Operationalizing the Requirements for AI Systems – Part I" (2022), at [https://openloop.org/reports/2022/11/Artificial\\_Intelligence\\_Act\\_A\\_Policy\\_Prototyping\\_Experiment\\_Operationalizing\\_Reqs\\_Part1.pdf](https://openloop.org/reports/2022/11/Artificial_Intelligence_Act_A_Policy_Prototyping_Experiment_Operationalizing_Reqs_Part1.pdf)
- 2 Given the importance of transparency as a requirement under the AIA and different aspects of transparency (e.g., transparency of the AI development process versus transparency of individual AI decisions) we wanted to follow up on this from the Open Loop Forum.
- 3 See <https://openloop.org/programs/open-loop-eu-ai-act-program/>
- 4 For the purposes of this report, we primarily focused on the results obtained from the providers as they are the primary recipients of the AIA requirements. However, we also incorporated relevant insights from the advisors to provide a comprehensive overview of the findings.
- 5 We used an ethnography application (dScout) to gather feedback from participants: [www.dscout.com](http://www.dscout.com)
- 6 The element 'use it appropriately' can refer grammatically to both the outcome of the AI system as well as the AI system itself. Our assumption is that 'use it appropriately' points to both the use of the output and an appropriate use of the AI system itself. That means that the AI system must be designed in such a way that the operation is sufficiently transparent for users to interpret the output (requirement 1) and that the AI system is accompanied with the right instructions and information such that users are capable of operating the AI system in the right manner (requirement 2).
- 7 Note that participants were split between the risk management and transparency deep dives.
- 8 Andrade, Norberto Nuno Gomes de, and Antonella Zarra. "Artificial Intelligence Act: A Policy Prototyping Experiment: Operationalizing the Requirements for AI Systems – Part I" (2022), at [https://openloop.org/reports/2022/11/Artificial\\_Intelligence\\_Act\\_A\\_Policy\\_Prototyping\\_Experiment\\_Operationalizing\\_Reqs\\_Part1.pdf](https://openloop.org/reports/2022/11/Artificial_Intelligence_Act_A_Policy_Prototyping_Experiment_Operationalizing_Reqs_Part1.pdf)
- 9 See overview of Open Loop programs deployed till this date at <https://openloop.org/lets-unlock/>
- 10 The element 'use it appropriately' can refer grammatically to both the outcome of the AI system as well as the AI system itself. Our assumption is that 'use it appropriately' points to both the use of the output and an appropriate use of the AI system itself. That means that the AI system must be designed in such a way that the operation is sufficiently transparent for users to interpret the output (requirement 1) and that the AI system is accompanied with the right instructions and information such that users are capable of operating the AI system in the right manner (requirement 2).
- 11 This specific aspect will be evaluated separately within the Open Loop EU AIA Program.
- 12 The set of results returned from a query.
- 13 Andrade, Norberto Nuno Gomes de, and Antonella Zarra. "Artificial Intelligence Act: A Policy Prototyping Experiment: Operationalizing the Requirements for AI Systems – Part I" (2022), at [https://openloop.org/reports/2022/11/Artificial\\_Intelligence\\_Act\\_A\\_Policy\\_Prototyping\\_Experiment\\_Operationalizing\\_Reqs\\_Part1.pdf](https://openloop.org/reports/2022/11/Artificial_Intelligence_Act_A_Policy_Prototyping_Experiment_Operationalizing_Reqs_Part1.pdf)
- 14 By project, we mean the deployment of a certain AI system/model.
- 15 The number of hours consist of hours needed to design these systems in such a way that human oversight is enabled (design costs), not actual costs of implementing oversight (performative costs).
- 16 MLOps stands for Machine learning Operations. It is the process of developing machine learning models, putting them into production and maintaining them.

- 17 Andrade, Norberto Nuno Gomes de, and Antonella Zarra. "Artificial Intelligence Act: A Policy Prototyping Experiment: Operationalizing the Requirements for AI Systems – Part I" (2022), at [https://openloop.org/reports/2022/11/Artificial\\_Intelligence\\_Act\\_A\\_Policy\\_Prototyping\\_Experiment\\_Operationalizing\\_Reqs\\_Part1.pdf](https://openloop.org/reports/2022/11/Artificial_Intelligence_Act_A_Policy_Prototyping_Experiment_Operationalizing_Reqs_Part1.pdf), p. 33.
- 18 Black box models are models that are not readily interpretable to humans, globally interpretable models on the other hand can be fully understood by humans. See: Molnar, C. (2022), Interpretable Machine Learning: A Guide for Making Black Box Models Explainable, 2nd edition via: <https://christophm.github.io/interpretable-ml-book/index.html>
- 19 Shapley Additive Explanations (SHAP) is an approach to explaining the outcomes of any machine learning model.
- 20 See also: Meta AI - System Cards, <https://ai.facebook.com/tools/system-cards/>
- 21 Andrade, Norberto Nuno Gomes de. "AI Transparency and Explainability - A Policy Prototyping Experiment" (2022), at: [https://openloop.org/wp-content/uploads/2022/08/OPENLOOP\\_asia\\_pacific\\_FullReport\\_2022.pdf](https://openloop.org/wp-content/uploads/2022/08/OPENLOOP_asia_pacific_FullReport_2022.pdf)
- 22 OpenLoop's AI Transparency and Explainability program in Singapore provides useful insights into how to build AI explainability for a range of use cases and stakeholders in a more holistic and comprehensive way. See: <https://openloop.org/programs/ai-transparency-explainability-singapore-2/>
- 23 As one participant observed, the AIA is very vague in some parts and very prescriptive in others: "Overall I would describe it as confusing. Some elements are overly prescriptive, while others too little." The Newsroom.
- 24 This was a specific issue mentioned in the context of the OLF: many providers used pre-trained models from third parties and could not describe them in the requisite level of detail, because they did not have access to that information.
- 25 This was a specific issue mentioned in the context of the OLF: providers feared that they would have to share confidential information.
- 26 Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. <https://doi.org/10.48550/ARXIV.1810.04805>
- 27 This is partly due to the fact that several of the participants' AI systems are still in the development stage or because their requested patents are still pending.
- 28 The expressed opinions are from the perspective of the provider and do not reflect the opinions of the product's users.
- 29 Feature drift occurs when a machine learning model's performance declines due to changes in the distribution of features in the data it is applied to.
- 30 Model performance in production refers to how well a model performs when it is used to make predictions or decisions in a real-world setting.
- 31 Model input/output distribution refers to the distribution of possible inputs and outputs that a model can handle.
- 32 Model training and re-training refers to the process of building and improving a model by adjusting its parameters based on data.
- 33 Model evaluation and testing refers to the process of measuring a model's performance on a dataset to see how well it generalizes to new data.

- 34 Hardware metrics refer to characteristics of a computer or other hardware that can affect a model's performance.
- 35 CI/CD pipelines for ML refer to the process of automatically building, testing, and deploying machine learning models.
- 36 Accuracy of predictions and accuracy in classification refers to how often a model's predictions or classifications are correct.
- 37 Data mismatch, data integrity, and data drift refer to issues that can arise when the data used to train a model does not accurately reflect the data the model will encounter in production, or when the data used to train a model changes over time.
- 38 Classification accuracy is a measure of how many predictions a model got correct.
- 39 Logarithmic loss is a measure of how confident a model is in its predictions.
- 40 A confusion matrix is a table showing the number of true positive, true negative, false positive, and false negative predictions.
- 41 The area under curve is a measure of model performance for a binary classifier.
- 42 The F1 score is a measure of a model's precision and recall.
- 43 The mean absolute error is a measure of the absolute difference between predicted and actual values. The mean squared error is a measure of the average squared difference between predicted and actual values.
- 44 Measuring AUC of model predictions against human-labeled test sets is a way to evaluate model performance.
- 45 Using related metrics that take different perspectives on the data and outcomes can provide a more comprehensive understanding of a model's strengths and weaknesses.
- 46 Model weakness is an aspect of a model that performs poorly or is prone to error.
- 47 Decision boundary detection is the process of identifying the points at which a model's predictions change.
- 48 Scenario testing is the process of evaluating a model's performance under different conditions or situations.
- 49 Perturbation testing is the process of evaluating a model's robustness by making small changes to the input data and observing the impact on the model's predictions.
- 50 Behavioral testing is the process of evaluating a model's performance by comparing its predictions to human-labeled data.
- 51 Data is partitioned by splitting the features and accuracy is calculated for each partition.
- 52 Data is partitioned and a separate model is trained for each partition with same parameters. Accuracy is calculated for each model.
- 53 The data distribution is shifted by applying a transformation to the features. Accuracy is calculated for various shifts.
- 54 The model is run repeatedly for some batch of data, each time some random set of neurons dropped. Accuracy is calculated for each case.
- 55 The CRISP-DM (Cross-Industry Standard Process for Data Mining) is a systematic approach to data mining and machine learning that involves six steps: business understanding, data understanding, data preparation, modeling, evaluation, and deployment.