cerre Centre on Regulation in Europe

EFFECTIVE IMPLEMENTATION OF REQUIREMENTS FOR HIGH-RISK AI SYSTEMS UNDER THE AI ACT: TRANSPARENCY AND APPROPRIATE ACCURACY

ISSUE PAPER

February 2025

Daniel Schnurr

Issue Paper

Effective Implementation of Requirements for High-Risk AI Systems Under the AI Act: Transparency and Appropriate Accuracy

Daniel Schnurr

February 2025

As provided for in CERRE's bylaws and procedural rules from its “Transparency & Independence Policy”, all CERRE research projects and reports are completed in accordance with the strictest academic independence.

The project, within the framework of which this report has been prepared, received the support and/or input of the following CERRE member organisations: Amazon, Booking.com, Microsoft, Mozilla, Tata Consultancy Services, Institut Belge des Services Postaux et des Télécommunications (IBPT), Hellenic Telecommunications and Post Commission (EETT). However, they bear no responsibility for the contents of this report. The views expressed in this CERRE report are attributable only to the authors in a personal capacity and not to any institution with which they are associated. In addition, they do not necessarily correspond either to those of CERRE, or of any sponsor or of members of CERRE.

info@cerre.eu – www.cerre.eu

Table of Contents

About CERRE

Providing high quality studies and dissemination activities, the Centre on Regulation in Europe (CERRE) is a not-for-profit think tank. It promotes robust and consistent regulation in Europe's network and digital industry and service sectors as well as in those impacted by the digital and energy transitions. CERRE's members are regulatory authorities and companies operating in these sectors, as well as universities.

CERRE's added value is based on:

- its original, multidisciplinary and cross-sector approach covering a variety of markets, e.g., energy, mobility, sustainability, tech, media, telecom, etc.;
- the widely acknowledged academic credentials and policy experience of its research team and associated staff members;
- its scientific independence and impartiality; and,
- the direct relevance and timeliness of its contributions to the policy and regulatory development process impacting network industry players and the markets for their goods and services.

CERRE's activities include contributions to the development of norms, standards, and policy recommendations related to the regulation of service providers, to the specification of market rules and to improvements in the management of infrastructure in a changing political, economic, technological, and social environment. CERRE's work also aims to clarify the respective roles of market operators, governments, and regulatory authorities, as well as contribute to the enhancement of those organisations' expertise in addressing regulatory issues of relevance to their activities.

About the Author

Daniel Schnurr is a CERRE Research Fellow and a Professor of Information Systems at the University of Regensburg, where he holds the Chair of Machine Learning and Uncertainty Quantification.

Previously, he led the Data Policies research group at the University of Passau. He received his Ph.D. in Information Systems from the Karlsruhe Institute of Technology in 2016, where he also completed his B.Sc. and M.Sc. in Information Engineering and Management. Daniel Schnurr has published in leading journals in Information Systems and Economics.

His current research focuses on the role of artificial intelligence for competition, privacy and data sharing in digital markets, as well as regulation of AI and the data economy.

Executive Summary

This Issue Paper takes a first step toward analysing and deriving recommendations for the efficient and effective operationalisation of selected requirements under the AI Act (AIA). The paper identifies open questions about the implementation of provisions for high-risk AI systems, evaluates potential approaches to address these questions in light of the underlying trade-offs, and proposes further steps to establish actionable guidance for providers and deployers of high-risk AI systems. The analysis focuses specifically on the transparency provisions in Art. 13 AIA and the provisions on appropriate accuracy in Art. 15 AIA for high-risk AI systems.

A central challenge to the implementation of the AIA lies in its broad scope. Given the wide range of technical approaches employed in high-risk AI systems and the diverse use cases to which they are applied, deriving general yet useful and appropriate criteria is inherently difficult. To address this challenge, the paper proposes elements and principles that could be established at the general level of the AIA to promote consistency across risk domains and applications. These recommendations include defining main categories of information to be provided under the transparency requirement and establishing a general principle for determining the appropriate level of accuracy of high-risk AI systems. Furthermore, the paper highlights requirements and specifications that demand more granular guidance, such as within specific risk domains, to ensure effectiveness and proportionality. For example, this includes criteria to determine whether the interpretability of a high-risk AI system's outputs requires using intrinsically interpretable models or post-hoc interpretation techniques, as well as guidance on what performance dimensions constitute relevant accuracy criteria for a given high-risk AI system.

Based on an analysis of the AIA provisions, the specific requirements, the technical state of the art in AI systems and methodologies, findings from academic literature, and the involved trade-offs, the following key recommendations are derived:

1. **Implementation of AIA transparency requirements can be operationalised by delineating three main categories of information**
   a) Information on the characteristics, capabilities and limitations of performance of a high-risk AI system.
   b) Documentation of potential risks, and known or foreseeable circumstances that negatively impact functional or non-functional characteristics of the AI system.
   c) Information and measures to enable interpretability of AI system outputs.

   Transparency about the characteristics, capabilities and limitations of performance of a high-risk AI system can generally also satisfy requirements for the interpretation of AI system outputs. To convey this information, the instructions for use of high-risk AI systems could build on the concept of model cards, which may be augmented to also support the consistent documentation of potential risks and relevant influencing factors. In addition, information about the integration of the AI model into the broader AI system, and how additional data processing at the system level contributes to the AI system's outputs, should be included.

Beyond this general principle, risk-domain-specific analyses may establish the need for intrinsically interpretable models or post-hoc explanation techniques to allow for the interpretability of AI system outputs in specific cases - but must account for the limitations of these techniques and the involved trade-offs with other functional and non-functional objectives.

2. **Establishing common practices through guidelines and standardisation for specific risk domains**
   Standardisation and the agreed-upon selection of criteria and metrics promote transparency and risk mitigation, as deployers can more easily assess and evaluate provided information. At the same time, this can help providers of AI systems by reducing uncertainty and compliance costs through the adoption of common practices. However, the broad scope of the AIA requirements makes it difficult to set tangible and actionable agreements and standards, while also accounting for the various specificities of the diverse use cases and associated technical approaches underlying AI systems. Hence, guidelines and standards may be developed for specific risk domains of high-risk AI systems to promote common practices for implementation. The classification of high-risk systems specified in Art. 6 (1) and Art. 6 (2) AIA may provide a high-level delineation of these different application domains, although more granular approaches could be necessary in broad domains.
   Guidelines or standards at the level of risk domains can help identify use cases where interpretability of AI system outputs is deemed essential and establish whether intrinsically interpretable models or post-hoc explanation techniques should be considered suitable measures in these specific cases (see Recommendation 1). This requires an analysis of the feasibility and adequacy of these methods in the given context, the trade-offs with other performance and non-performance goals, and their effectiveness in mitigating the specific risks. Additionally, potential challenges to interpretability arising from the interplay between AI models and the system into which they are integrated should be considered.
   Furthermore, establishing common practices for specific risk domains can promote the effective implementation of transparency and accuracy requirements under the AIA by providing guidance on relevant accuracy criteria, metrics, testing procedures, and test sets. Such common practices can complement and support contractual agreements between actors along the AI value chain. This may include specifying commonly accepted templates on these different elements for high-risk providers to choose from for their use cases. In addition, standardised test datasets and testing procedures may be developed with regard to specific use cases and conditions in a particular risk domain. To this end, domain-specific standards should build on established or emerging standards for the transparency and accuracy of AI systems to reduce compliance costs and support global harmonisation.

3. **Timely transparency through post-market monitoring of risks and effective feedback mechanisms across the AI value chain**
   The analysis of the AIA transparency requirements and their intended purposes highlights the importance of post-market monitoring of risks and effective feedback mechanisms between deployers and providers of high-risk AI systems to ensure timely transparency about risks and to

facilitate cooperative approaches to risk mitigation between providers and deployers of high-risk AI systems.

Such feedback mechanisms should complement mandatory reporting obligations for serious incidents, which must be reported to the market surveillance authorities. In contrast, the envisioned feedback mechanisms for post-market monitoring aim to foster collaboration between providers and deployers to anticipate and mitigate emerging risks, consider limitations of performance and possible remedies that arise with respect to context-specific applications, or identify changes in the inputs and environments that could affect the performance of the AI system. The information gathered through such feedback mechanisms should then be dynamically incorporated into the instructions for use of the concerned high-risk AI system and made accessible to all deployers to promote timely transparency. In turn, effective post-market monitoring of risks can alleviate the burden of ex-ante risk identification and evaluation, which is often constrained by the complex, uncertain, and dynamic contexts of many high-risk AI systems.

4. **Human accuracy and technical state of the art as a general benchmark for the appropriate accuracy of high-risk AI systems**

As a general principle across risk domains, appropriate accuracy should be determined based on the state of the art of commonly available alternatives performing the same task. In many contexts where AI systems are expected to operate, the task has previously been performed by either a human or a non-AI technical system. Consequently, a lower bound for the appropriate accuracy of a high-risk AI system is determined by the average accuracy achieved by humans or, if a non-AI technical system that is superior to human accuracy is commonly applied to the task, by the accuracy of that system.

If other AI systems are already applied to the same task and considered state of the art, these systems should be included in the comparison of relevant alternatives. However, it is then important not to define the state of the art too narrowly, as this could undermine predictability, competition, and innovation. Therefore, to qualify as state of the art, the corresponding level of accuracy should have been established over a reasonable period of time and should not be considered exceptional.

Compliance with this proposed principle could establish a presumption of conformity. As an exception, high-risk AI systems may be permitted to operate even if they fail to meet the general principle, but only if the benefits provided by such systems are deemed sufficiently important from a societal perspective. In such cases, the burden of proof should be placed on the provider to demonstrate that the benefits of the high-risk AI system outweigh the risks associated with its lower accuracy. To promote innovation, the AI Office or national competent authorities may develop procedures or tools to support these assessments, thereby offering legal certainty to providers.

Whenever the accuracy for a given task is not measurable using accepted criteria and metrics, a possible benchmark for appropriate accuracy could be that the accuracy of the high-risk AI system is sufficient to avoid causing risks to the intended group of users or a specific subgroup of these users. As this benchmark requires significantly more consideration of system-specific and use case-specific contexts, evaluations would need to be conducted on a case-by-case basis.

1 Introduction

The obligations under the European AI Act (AIA)[1], especially those related to high-risk AI systems, specify a broad range of technical requirements for AI systems, their quality, data inputs, documentation, and transparency. As the regulation relies mostly on abstract goals and general principles to define these requirements, there remain open questions about their implementation in practice. This flexibility can be viewed as essential for ensuring technology neutrality and for accommodating the broad scope of the AI Act in terms of diverse application domains and rapidly evolving AI technologies as well as international standards and governance frameworks. At the same time, the ambiguity associated with this flexibility can lead to uncertainty and inconsistencies in implementation, posing a risk of stifling innovation and increasing costs for providers and deployers of AI systems in areas where AI has the potential to generate significant welfare benefits.

Therefore, this Issue Paper aims to take a first step toward analysing and deriving recommendations on how selected specific requirements under the AI Act can be operationalised efficiently and effectively from a practical perspective. To this end, the Issue Paper identifies open questions for implementation of selected provisions on high-risk AI systems, analyses potential options for how to address these questions in the light of the underlying trade-offs, and recommends further steps for establishing more actionable guidance for providers and deployers of high-risk AI systems. In this vein, the Issue Paper also aims to contribute to ongoing initiatives to establish common practices and harmonised standards for the AIA.[2] The Issue Paper specifically focuses on the transparency provisions in Art. 13 AIA and the provisions on appropriate accuracy in Art. 15 AIA for high-risk AI systems. Under the AIA, an AI system is classified as high-risk based on the criteria and classification rules set out in Art. 6 AIA. As the high-risk classification draws on existing product safety legislation[3] and a predefined list of diverse risk domains[4], the AIA requirements for high-risk AI systems apply horizontally across a broad and heterogeneous set of use cases and application domains.

This diversity presents a central challenge to the implementation of the AIA requirements. On the one hand, requirements should be implemented consistently and where feasible in a standardised manner across application areas to minimise costs for compliance and facilitate effective enforcement. On the other hand, the specificities of different application areas and use cases make it difficult to set absolute thresholds or specific metrics for satisfying requirements. Additionally, mechanisms that could be adequate implementations of requirements in one area of application may be difficult or even unfeasible to implement in other areas. For example, the level of accuracy that is deemed appropriate for a high-risk AI system and is thus considered acceptable can vary significantly between applications and contexts.

[1] Regulation 2024/1689 of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), http://data.europa.eu/eli/reg/2024/1689/oj [hereinafter AIA].

[2] See European Commission (2023). Commission Implementing Decision on a standardisation request to the European Committee for Standardisation and the European Committee for Electrotechnical Standardisation in support of Union policy on artificial intelligence. Register of Commission - Documents C(2023)3215. Available at https://ec.europa.eu/transparency/documents-register/detail?ref=C(2023)3215&lang=en

[3] Art. 6 (1) and Annex I AIA.

[4] Art. 6 (2) and Annex III AIA.

Furthermore, there exist interdependencies and possibly tensions between different requirements and their ability to address the risks of AI systems. For example, increased interpretability of AI models as a means to improve the transparency of AI systems regularly comes at the cost of decreased prediction accuracy. As the superior accuracy of black-box models often stems from their ability to capture the complex relationships and patterns of the underlying phenomena, capturing the same phenomena by simpler, human-interpretable models requires a reduction of model complexity, typically at the expense of prediction accuracy.[5] Depending on the precise use case and types of the associated risks, different requirements and implementation approaches may therefore be prioritised. The diversity in use cases also implies that the technical foundations of high-risk AI systems may significantly differ, which reinforces challenges for consistent and efficient implementation. For example, depending on whether AI systems process tabular data, images, text or video, the performance metrics as well as potential risks and mitigation strategies may differ significantly. This is further complicated by the ongoing technical progress and short innovation cycles in these technologies, which may quickly change the characteristics and inner workings of AI systems in specific application areas.

A first step toward effective and efficient implementation, therefore, calls for an analysis of which requirements call for domain-specific guidance to support implementation and, where feasible, this can be supported by generalisable principles that apply horizontally across risk domains and areas of application. While an exhaustive analysis of the requirements for high-risk AI systems under the AIA in this regard is beyond the scope of this paper, the Issue Paper aims to inform this broader discussion by providing an exemplary analysis of the AIA transparency provisions and the provisions on appropriate accuracy. In addition, this Issue Paper makes specific recommendations on how to operationalise transparency and interpretability requirements for high-risk AI systems based on an analysis of the AIA provisions, the involved trade-offs and the academic literature on explainability and interpretability of AI.

These recommendations suggest that transparency about the characteristics, capabilities and limitations of performance of a high-risk AI system can satisfy requirements on the interpretation of AI system outputs in general. Risk-domain-specific analyses may establish the need for intrinsically interpretable models or post-hoc explanations to ensure interpretability of AI system outputs in specific cases but must account for the limitations of these techniques and the involved trade-offs with other functional and non-functional objectives. To promote effective and efficient implementation, such additional interpretability requirements should be established through guidelines or standardisation at the level of the particular risk domain with regard to specific use cases and contexts of high-risk AI systems. Guidelines or standardisation for individual risk domains may further set out more specific requirements on performance metrics, testing procedures and possibly standardised datasets to support effective transparency reporting. While this Issue Paper concludes that implementation approaches should be developed with regard to specific risk domain levels to establish actionable common practices, there remains the need for the alignment of these practices within the general framework of the AIA to ensure horizontal consistency. Finally, this Issue Paper highlights the importance of post-market monitoring of

[5] See Section 3.3 for a more detailed discussion of the general trade-off between interpretability and accuracy of AI systems.

risks and effective feedback mechanisms between deployers and providers of high-risk AI systems to ensure timely transparency about risks and risk mitigation.

Regarding the accuracy requirements for high-risk AI systems, this Issue Paper discusses the challenges in devising standards for accuracy specifications and identifying what could constitute a benchmark for appropriate quality. As a general principle across risk domains, appropriate accuracy should be determined based on the state of the art of commonly available alternatives performing the same task. In several contexts where AI systems are expected to operate, the task has previously been performed by either a human or a non-AI technical system. Consequently, a lower bound for the appropriate accuracy of a high-risk AI system is established by the average accuracy achieved by humans or, if a non-AI technical system that is superior to human accuracy is commonly applied to the task, by the accuracy of that system. If other AI systems are already applied to the same task and considered state of the art, the comparison of relevant alternatives should include these AI systems. However, it is then important to not define the state of the art too narrowly to safeguard predictability, competition, and innovation.

Overall, this Issue Paper thus contributes to selected substantive elements of the AIA, while not addressing institutional questions of implementation and enforcement. Because the institutions responsible for overseeing the implementation of the technical requirements are equally diverse and the institutional framework can be anticipated to be highly complex, this will add another layer of challenges for efficient and effective implementation. A first assessment of the institutional governance framework and major procedural issues are covered in the companion CERRE Issue Paper.[6]

The remainder of the Issue Paper is organised as follows: First, Section 2 outlines the main goals of the AIA and its approach to high-risk AI systems, which draws on the requirements of trustworthy AI. Section 3 addresses the implementation of transparency requirements for high-risk AI systems according to Art. 13 AIA. Section 4 discusses the AIA's provisions on accuracy in Art. 15 AIA and focuses on the question of how to define a general benchmark for appropriate accuracy. Section 5 concludes by summarising the main insights and recommendations derived from the analysis.

[6] Larouche, P. (2025). Legal framework for an effective implementation of the AI Act. CERRE Issue Paper. Available here.

2 AIA Requirements for High-Risk AI Systems

The AIA is intended to "promote the uptake of human-centric and trustworthy artificial intelligence (AI) while ensuring a high level of protection of health, safety, fundamental rights […], to protect against the harmful effects of AI systems in the Union, and to support innovation".[7] In this context, it is explicitly mentioned that requirements for AI systems in the European Union need to be consistent and harmonised to avoid fragmentation of the internal market and to reduce uncertainties for operators of AI systems.[8]

To balance the goals of AI safety and innovation, the AIA adopts a risk-based approach, whereby stricter requirements are imposed on AI systems with higher anticipated risks. This approach is consistent with the general principle of proportionality, which is prominently referenced in many provisions of the AIA. While Chapter II of the AIA prohibits certain AI practices deemed to pose unacceptable risks,[9] Chapter III defines classification rules for high-risk AI systems and imposes a set of requirements for these systems. Following the model of the New Legislative Framework[10], the enforcement procedure established by the AIA includes a pre-market conformity assessment, post-market monitoring and surveillance, as well as fines and penalties for significant breaches.[11]

Art. 6 AIA classifies high-risk AI systems based on their application in risk domains defined either by Union harmonisation legislation covering products that entail safety risks and foresees a conformity assessment procedure, or by a predefined list of high-risk areas outlined in Annex III AIA.[12] Consequently, requirements on high-risk AI systems cover diverse risk domains such as machinery, toys, medical devices, education and vocational training, employment, and critical infrastructures. Additional use cases may be incorporated into Annex III through delegated acts adopted by the Commission.[13]

As a central part of the enforcement procedure, Art. 8 and Art. 9 AIA require that high-risk AI systems are accompanied by a risk management system, which "shall be understood as a continuous iterative process planned and run throughout the entire lifecycle of a high-risk AI system, requiring regular systematic review and updating".[14] The two main purposes of the risk management system are (i) the identification, estimation, and analysis of risks that may emerge when the AI-system is used and (ii) the adoption of appropriate and targeted risk management measures to mitigate identified risks that can emerge when the AI system is used in accordance with its intended purpose.[15] In this context, risk management measures shall consider the effects and possible interaction with the combined application of other requirements set out in Section III AIA.[16] In addition, Art. 9 (5) (c) explicitly references the transparency

[7] Recital 1 AIA

[8] Recital 2 AIA; Under the AIA, operator is a catch-all term to include providers, product manufacturers, deployers, authorised representatives, importers or distributors as defined by Art. 3 (8) AIA.

[9] Chapter II AIA

[10] See for an overview: European Commission. (n.d.). New legislative framework. Available at https://single-market-economy.ec.europa.eu/single-market/goods/new-legislative-framework_en

[11] Larouche, P. (2025). Legal framework for an effective implementation of the AI Act. CERRE Issue Paper.

[12] Art. 6 (1) and (2) AIA. See also Recitals 47, 48 and 50 AIA.

[13] Art. 7 and Art. 97 AIA.

[14] Art. 9 (2) AIA

[15] Art. 9 (2)(a) to (d) AIA

[16] Art. 9 (4) AIA.

requirements outlined in Art. 13 AIA, emphasising their necessity for identifying the most appropriate risk management measures.

Section 2 of Chapter III of the AIA establishes several mandatory requirements intended to mitigate the risks from AI systems such that their outputs do not pose unacceptable risks to important public interests and to ensure a high level of trustworthiness.[17] Measures to comply with these requirements should be proportionate and effective to meet the objectives of the AIA and should take into account the intended purpose and context of use of the AI system, as well as the generally acknowledged state of the art on AI.[18] This suggests that while all requirements set out in Section 2 of Chapter III of the AIA are mandatory to be addressed for any high-risk AI system,[19] the implementation of specific measures and necessary thresholds to fulfil the requirements can vary across contexts and use cases as well as within contexts depending on the anticipated level of risk. Therefore, the level of risk is determined by the severity of a possible harm and its probability of occurrence.[20] As such, the AIA also foresees that requirements should be applied according to the risk-management system to be established by the provider.[21] In general, the categorisation of risks presents a major challenge for adequate and proportionate implementation, as the risk levels of high-risk AI Systems may still vary significantly. In addition, the risk level associated with an AI system is not static but can change dynamically.[22]

The requirements applying to high-risk AI systems as established by Section 2 of Chapter III of the AIA include the following provisions on:

- Data and data governance (Art. 10 AIA)
- Technical documentation (Art. 11 AIA)
- Record-keeping (Art. 12 AIA)
- Transparency and provision of information to deployers (Art. 13 AIA)
- Human oversight (Art. 14 AIA)
- Accuracy, robustness and cybersecurity (Art. 15 AIA)

These requirements draw heavily on generally acknowledged properties and principles of *trustworthy AI*. Trustworthy AI is generally defined by a broad range of characteristics of AI systems including being “valid and reliable, safe, secure and resilient, accountable and transparent, explainable and interpretable, privacy-enhanced, and fair with harmful bias managed”.[23] Ensuring the characteristics of trustworthy AI is intended to reduce the potential negative risks associated with AI systems and to promote adoption and acceptance by increasing the trust that humans can place in those systems. With the success of AI

[17] Recital 46 and Recital 64 AIA.

[18] Recital 64 AIA.

[19] Recital 66 AIA states that the requirements laid out in Section 2 of Chapter III “are necessary to effectively mitigate the risks for health, safety and fundamental rights” and that “no other less trade restrictive measures are reasonably available.”

[20] Recital 52 AIA.

[21] Recital 64 AIA.

[22] In consequence, mechanisms and institutions intended to support the implementation of requirements for high-risk AI systems should allow for dynamic adjustments to risk classifications and assist operators of AI systems in assessing their system’s risk level.

[23] NIST (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). Available at https://doi.org/10.6028/NIST.AI.100-1, p. 12.

systems in a wide range of use cases and application contexts, the field of trustworthy AI has become a rapidly growing area of research aimed at defining criteria for trustworthy AI systems, conceptualising AI governance frameworks, developing new methods for ensuring properties such as interpretability and explainability, and exploring the implementation and empirical evaluation procedures of trustworthy AI concepts.

In 2020, the High-Level Expert Group on Artificial Intelligence (HLEG AI) appointed by the European Commission presented their own Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment based on the previously drafted Ethics Guidelines for Trustworthy AI.[24] The requirements for high-risk AI systems under the AIA share several of the criteria developed by the HLEG AI. The Ethics Guidelines are further referenced by the AIA as a complementary set of rules that operators may voluntarily adhere to in addition to the binding rules of the AIA.[25]

The need for new concepts and measures to ensure trustworthy AI beyond conventional risk mitigation strategies for traditional software or information-based systems is typically justified by the unique properties of AI systems.[26] In particular, AI systems (especially the underlying AI models) and the application contexts where they are deployed are often highly complex, while system outputs are usually probabilistic, making it difficult to detect and respond to failures. Furthermore, AI systems are often pre-trained on data. This data may change over time or may only represent a subset of the data that systems will face when deployed, thus leading to risks of inaccurate generalisations or unnoticed distribution shifts. As AI systems mimic human behaviour, they may further act as socio-technical systems. In such use cases, AI systems are influenced by societal dynamics and human behaviour, either through the training procedures or during inference when interacting with human users or other AI systems. Consequently, AI systems may be prone to reproducing undesired behaviour and outcomes, possibly at a much larger scale and without clear traceability or accountability for how these outcomes emerge. Moreover, due to their higher level of autonomy, AI systems may produce emergent behaviour after deployment, influenced by the input of human users or other AI systems, for which the outcomes are difficult to fully anticipate and characterise ex-ante.

These potential risks of AI systems will typically materialise during the inference phase, when the AI system is deployed in a specific context for a particular use. While some risk management measures can be taken by the deployer of the AI system, other measures require access to the AI system or need to be taken already at the development, training or testing stage of the AI system and thus need to be implemented by the provider of the AI system. There may be a further distinction of roles along the AI value chain in cases where providers of an AI system rely on AI models, which are developed and provided by other providers, as inputs for their own systems. In particular, this is typically the case for general-

[24] High-Level Expert Group on Artificial Intelligence (HLEG AI). 2020. Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment. Available at https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment; HLEG AI (2019). Ethics guidelines for trustworthy AI. Available at https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

[25] Recital 7 and Recital 27 AIA.

[26] NIST (2023).

purpose AI (GPAI) models, which can be used in a broad range of application contexts and use cases and are thus procured and integrated by providers as inputs for more specialised AI systems.

This value chain does not only imply that risk management measures can be taken at different levels, and that risk mitigation frequently involves shared responsibilities, but also that effective risk mitigation will require a continuous process that is not only concerned with pre-deployment conformity but also takes into account the monitoring of risks after deployment. Consequently, effective risk mitigation and implementation of the AIA requirements for high-risk systems also call for coordination and information exchange between the different actors across the AI value chain. Most notably, relevant problems and failures of AI systems will regularly first be observed by the deployer of an AI system. This information must then reach the provider of the AI system to allow for effective risk mitigation at the upstream level. In turn, updates to the AI system can be made available by the provider of the AI system but usually need to be installed by the deployer of the system. In addition, the provider of an AI system may rely on information obtained from deployers through specific requests to identify and evaluate potential risks.

The AIA acknowledges the vertical AI value chain and considers distinct roles and responsibilities for risk mitigation.[27] For high-risk AI systems, the AIA puts the burden of compliance with the requirements set out in Section 2 of Chapter III AIA on the provider of the AI system.[28] However, according to Art. 25 AIA, there may be more than one provider of an AI system, specifically if a distributor, importer, deployer or a third-party puts their name or trademark on a high-risk system, substantially modifies a high-risk AI system, or modifies the intended purpose of an AI system such that this system becomes a high-risk AI system.

The provider of a high-risk AI system must ensure that the high-risk AI system undergoes the relevant conformity assessment procedure, as detailed in Art. 43 AIA, and must take necessary corrective actions according to Art. 20 AIA if a high-risk system is found not in conformity with the AIA requirements after being placed on the market.[29] To ensure compliance, providers of high-risk AI systems must implement a quality management system that, among other elements, documents the risk management system as referred to in Art. 9 AIA as well as the technical specifications and means to be used to ensure that the requirements set out in Section 2 of Chapter III AIA are met.[30] Where available and applicable these technical specifications should include standards to ensure compliance with the requirements for high-risk AI systems. In addition, providers of high-risk AI systems are required to implement a post-market monitoring system according to Art. 72 AIA to ensure continuous compliance with the requirements set out in Section 2 of Chapter III AIA after a high-risk system is placed on the market or put into service.[31] As a further principle of proportionality, the implementation of the quality management system and its documentation should be proportionate to the size of the provider's organisation.[32] However, at the same

[27] See Section 3 of Chapter III AIA and Art. 25 in particular.
[28] Art. 16 (a) AIA.
[29] Art. 16 (f) and Art. 16 (j) AIA.
[30] Art. 17 (1)(e) and (g) AIA.
[31] Art. 17 (1); Art. 3 (25) AIA.
[32] Art. 17 (2) AIA.

time, it is required that "in any event" sufficient measures must be taken to ensure compliance with the requirements for high-risk AI systems.

Besides obligations for importers and distributors of high-risk AI systems,[33] the AIA also imposes obligations on the deployers, which mainly refer to measures to ensure the use of high-risk AI systems in accordance with the instructions for use issued by the provider of the high-risk AI system as well as to ensure human oversight of the system as required by Art. 14 AIA.[34] Furthermore, the deployer is responsible for keeping logs that are automatically generated by the high-risk AI system and are under its control.[35] If a high-risk system is deployed in the workplace, the deployer must ensure transparency for end users about the use of the system by informing affected workers and their representatives.[36] Where applicable, deployers are further responsible for carrying out a fundamental rights impact assessment and a data protection impact assessment before deploying the high-risk AI system.[37] In addition, deployers are required to monitor the operation of the high-risk AI systems and provide relevant information to providers in the context of their post-market monitoring system, in cases where the AI system is considered to present a risk according to Art. 79 AIA, and when they identify a serious incident as defined by Art. 73 AIA.[38]

Next to obligations for providers and deployers of high-risk AI systems, the AIA lays out specific obligations for GPAI model providers in Art. 53 AIA and additionally in Art. 55 AIA in the case of GPAI models with systemic risk. These obligations are outside of the scope of this paper's analysis, as the remainder of this Issue Paper will focus on high-risk AI systems and specifically examine transparency requirements and the requirement of appropriate accuracy in the following two sections.

[33] See Art. 23 and Art. 24 AIA.
[34] Art. 26 (1) and (2) AIA.
[35] Art. 26 (6) AIA.
[36] Art. 26 (7) AIA.
[37] Art. 27 and Art. 26 (9) AIA.
[38] Art. 26 (5) AIA.

3 Transparency Requirements and Interpretability of AI System Outputs

Art. 13 AIA is concerned with the transparency and interpretability of outputs of a high-risk AI system from the perspective of the deployers of these systems. These transparency requirements are imposed in addition to the requirements of technical documentation, as defined by Art. 11 AIA and specified in more detail in Annex IV AIA, which must allow national competent authorities and notified bodies to assess compliance of the AI system with the requirements set out in Section 2 of Chapter III AIA. Hence, Art. 13 AIA addresses potential information asymmetries between the provider and the deployer of an AI system to support effective risk mitigation and promote the trustworthiness of high-risk AI systems.

Art. 13 (1) AIA states that "[h]igh-risk AI systems shall be designed and developed in such a way as to ensure that their operation is sufficiently transparent to enable deployers to interpret a system's output and use it appropriately." Thereby, an appropriate type and degree of transparency shall enable the provider and the deployer to achieve compliance with the relevant obligations set out in Section 3 of Chapter III AIA. The main instrument foreseen by the AIA to achieve transparency and interpretability of high-risk AI systems are the instructions for use that providers should make available to deployers. Next to requirements of conciseness, completeness, correctness, clarity, relevance, accessibility and comprehensibility, Art. 13 (3) establishes a list of substantiative elements that the instructions for use should cover. In part, these elements are a subset of the information required for the technical documentation of a high-risk AI system as outlined in Art. 11 and Annex IV AIA. The elements referenced in Art. 13 (3) AIA mainly relate to informing the deployer about the purpose, characteristics, capabilities, limitations, and risks of the high-risk AI systems.[39] In addition to these elements, Art. 13 (3) references further requirements on enabling deployers to interpret the output of the high-risk AI system and to use it appropriately. Art. 13 (3) (d) AIA further refers to the interpretability of system outputs in the context of human oversight measures as required by Art. 14 AIA.

The transparency obligations set out in Art. 13 AIA explicitly target the deployers of high-risk AI systems and do not make reference to end users of these systems.[40] Yet, the references to the interpretability of system outputs imply that transparency requirements should facilitate the understanding and use of these systems by *human users* acting on behalf of the deployer.[41] Furthermore, the transparency obligations refer to the human oversight measures in Art. 14 AIA, which require that high-risk AI systems must be designed and developed such that they can be effectively overseen by natural persons.[42]

[39] See also Recital 72 AIA.

[40] Art. 86 AIA specifies a right to explanation for any affected person (i.e., individuals that may represent end users of an AI system or may be affected by its decision) subject to a decision which is taken by the deployer on the basis of the output from a high-risk AI system listed in Annex III (with the exception of point 2 thereof) and which produces legal or similarly significant effects. However, this right refers to an explanation of the role of the AI system in the decision-making procedure rather than an explanation of the AI system's output. Furthermore, Art. 50 AIA imposes transparency obligations for AI systems intended to interact directly with natural persons, requiring disclosure to end users about the AI identity of their interaction counterpart.

[41] Art. 13 (1) and Art 13 (3)(b)(vii) AIA.

[42] Art 13 (3)(d) and Art. 14 (1) AIA.

The transparency requirements are intended to support the design of high-risk AI systems such that they enable deployers "to understand how the AI system works, evaluate its functionality, and comprehend its strengths and limitations".[43] This should assist deployers in the use of the system and in making informed decisions, especially about the choice of a system and the choice of appropriate applications. According to the AIA, the need for mandatory transparency requirements is justified by concerns about the opacity and complexity of certain AI systems.[44] As many modern AI systems rely on complex machine learning models that represent black boxes in terms of how they transform inputs into outputs, transparency of these systems and interpretability of their outputs have been widely acknowledged challenges to establishing the trustworthiness of these systems.

3.1 Concepts of Transparency, Interpretability and Explainability

The academic literature on trustworthy AI as well as standardisation initiatives on AI risk management typically differentiate between concepts of transparency, explainability, and interpretability. However, despite a quickly growing body of research, there is no clear consensus on the precise meaning of each of these concepts, as they are used in different variations and sometimes interchangeably.

The US National Institute of Standards and Technology (NIST) states that transparency addresses "what happened" in a system, explainability refers to "how" a decision was made in a system, and interpretability can answer "why" a decision was made by a system to the user as well as its meaning and context.[45] While transparency, explainability, and interpretability are considered distinct characteristics, they are meant to support each other in assisting humans who operate, oversee or use an AI system to gain deeper insights into the functionality and trustworthiness of an AI system and its outputs.

Arrieta et al. define interpretability as the "ability to explain or to provide the meaning in understandable terms to a human".[46] In this context, understandability refers to the degree to which a human can understand the functions and the decisions made by an AI model, as the core component of an AI system. Explainability is further defined "as the details and reasons a model gives to make its functioning clear or easy to understand [, given a certain audience]".[47] In addition, an AI model is considered to be transparent "if by itself it is understandable".

Notably, the term *Explainable AI* is now used to describe a distinct and quickly growing field of research concerned with methods and processes that enable human users to understand, appropriately trust, and

[43] Recital 72 AIA.

[44] Recital 72 AIA.

[45] NIST (2023).

[46] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82-115, p. 85.

[47] Arrieta et al. (2020), p. 85.

produce more explainable AI models.[48] Closely related is the field of *Interpretable Machine Learning,* which explores methods and models that aim to make the behaviour and outputs of machine learning systems understandable to humans.[49]

Although there is no uniform terminology and taxonomy of the concepts of transparency, interpretability, and explainability in the context of AI systems, there is general agreement that these systems exhibit varying degrees of interpretability (referring to humans' understanding of the AI system and its outputs) and that interpretability can be promoted through different mechanisms and explainability methods. The feasible degree and the methods to achieve interpretability thereby crucially depend on the AI model employed in an AI system. Furthermore, plain "algorithmic transparency" of these AI models[50], such as disclosing the numerical values of learned technical parameters will often not be conducive to promoting human understanding of the AI system or its outputs, especially for complex machine learning models. For example, a trained large-language model (LLM) can be completely specified by the numerical values for its billions of parameters, but knowing these parameters will do little to help humans to get a better understanding of the AI model, the associated AI system and the generated outputs.

To operationalise the requirements set out in Art. 13 AIA, it is constructive to distinguish between two general levels of interpretability:

1. **Transparency or interpretability in a broader sense** refers to humans' understanding of the general workings and principles of the AI system. For example, NIST defines the transparency of an AI system as "the extent to which information about an AI system and its outputs is available to individuals interacting with such a system."[51] Transparency helps humans to understand the general logic of the methods that determine how an AI system will transform inputs into outputs. For example, knowing which class of machine learning models is used in a system can facilitate humans' understanding according to which methods and principles the AI system optimises its outputs. It can further allow humans to anticipate general capabilities and shortcomings of the AI system based on the knowledge about the properties of the used machine learning model. Having access to more detailed information about model training or the hyperparameters of the model or how the model was integrated into the system can allow for further inferences about the system's capabilities and its limitations. According to this understanding, interpretability in a broader sense also includes the "representation of the mechanisms underlying AI systems' operation", which is subsumed under the criterion of explainability by the NIST AI RMF taxonomy.[52]

[48] DARPA (2016). Broad Agency Announcement**.** Explainable Artificial Intelligence (XAI). DARPA-BAA-16-53. Available at https://www.darpa.mil/attachments/DARPA-BAA-16-53.pdf; Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., ... & Ranjan, R. (2023). Explainable AI (XAI): Core ideas, techniques, and solutions. *ACM Computing Surveys*, *55*(9), 1-33.

[49] Molnar, C. (2024). *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable.* Available at https://christophm.github.io/interpretable-ml-book/; Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, *116*(44), 22071-22080.

[50] The concept of "algorithmic transparency" should not be confused with the concept of "transparent models" which is sometimes used interchangeably with the intrinsically interpretable models (see also Section 3.3).

[51] NIST (2023), p. 15.

[52] NIST (2023), p. 16.

2. **Interpretability in the narrow sense** refers to humans' understanding of why an AI system produces a certain output.[53] Therefore, interpretability in the narrow sense can for example help humans to understand what properties of an instance (i.e., which features of the input data) would need to be changed to arrive at a different output of the AI system (for example, the classification of an input instance into another output class or the prediction of a lower or higher numerical output value).[54] In general, there are two approaches to achieve such interpretability for AI systems in the narrow sense. First, the use of intrinsically interpretable models (also called transparent models), which are "inherently self-explanatory and provide an immediate human-readable interpretation about how they transform certain inputs into outputs due to their structure."[55] Such self-explanatory models can be achieved by incorporating interpretability directly into the model structure.[56] For example, the model weights of a logistic regression model or the rules and structure of a decision tree can be readily interpreted by humans. Second, post-hoc interpretability methods can be applied to black-box models to generate explanations for their outputs. Such explanations may be provided on a local basis, that is, for each individual output of the model, or on a global basis, that is, for the overall (average) behaviour of the model (see also Section 3.3 and specifically p. 30 for further details).

3.2 AIA Transparency Requirements

Art. 13 AIA includes several requirements that aim to promote the transparency of high-risk AI systems and their underlying AI model (that is, interpretability in the broader sense). In this sense, the AIA aligns with the NIST AI RMF, which suggests that risks from a lack of explainability can be mitigated by describing how AI systems function.[57] Additionally, transparency measures are suggested to increase confidence in an AI system by promoting higher levels of understanding. The required information that providers shall include in the instructions for use to promote transparency can be viewed as a subset of the information required for the technical documentation of a high-risk AI.

Doshi-Velez and Kim argue that, on a fundamental level, the need for interpretability in the context of AI stems from an incompleteness in the problem formalisation that the AI system is supposed to address.[58] For example, with respect to fairness or other fundamental rights, the actual goals may be too abstract to be completely encoded in the system.[59] In many other cases, tasks may be too complex to enumerate all

[53] Here, interpretability in the narrow sense aligns with the concept of interpretability as defined by the NIST AI RMF (NIST 2023).

[54] This is the key idea behind the concept of *counterfactual explanations*, a popular approach to explain outputs of AI models that reveal what should have been different in an instance to observe a diverse outcome. See, for example, Guidotti, R. (2024). Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery*, *38*(5), 2770-2824.

[55] Bauer, K., Hinz, O., van der Aalst, W., & Weinhardt, C. (2021). Expl(AI)n it to me – Explainable AI and Information Systems Research. *Business & Information Systems Engineering*, *63*, 79-82.

[56] Du, M., Liu, N., & Hu, X. (2019). Techniques for interpretable machine learning. *Communications of the ACM, 63*(1), 68-77.

[57] NIST (2023), p. 16.

[58] Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. Available at https://doi.org/10.48550/arXiv.1702.08608.

[59] This abstraction is difficult, if not impossible, to resolve, as fairness metrics are numerous and often mutually exclusive. Consequently, enforcing a specific metric to define fairness carries the risk of inadvertently promoting other biases. See, for example, Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. Available at https://doi.org/10.48550/arXiv.1609.05807.

possible inputs and explore all boundaries wherein the AI system performs well. Interpretability through increased transparency is then one approach to "ensure that effects of gaps in problem formalisation" can become visible to deployers and users of the AI system. Similarly, Panigutti et al. argue that the provision of relevant documentation and information about a high-risk AI system (such as on the intended purpose of the system, potential edge cases or failure scenarios, or appropriate organisational measures to mitigate risks) can be an effective measure to achieve transparency and promote interpretability for deployers.[60]

While the AIA specifies the type of information that should be provided to deployers (as discussed below), the level of detail of this information is not explicitly referenced. In principle, the information should be provided in sufficient detail and completeness to assist deployers in using the AI system correctly and making informed decisions, to allow them to interpret the system's outputs, and to fulfil their compliance obligations set out in Art. 26 AIA.[61] While this remains a highly abstract benchmark without further contextualisation in the specific use cases of a high-risk AI system, it is informative that the AIA requires providers to share only selected information from the technical documentation with deployers. This reflects the different purposes and recipients of the technical documentation according to Art. 26 AIA and the transparency requirements according to Art. 13 AIA. Furthermore, providing selected and aggregate information promotes conciseness and can contribute to better accessibility and understandability, as required by Art. 13 (1). Aggregated information can further help to protect legitimate interests of the providers of AI systems in cases where too detailed information (for instance, about model characteristics or testing data) could reveal trade secrets or confidential information or compromise intellectual property rights.[62] However, to promote effective transparency, aggregation of information must maintain sufficient detail to allow the deployer to infer meaningful insights.

In addition to the transparency requirements in Art. 13 AIA, Art. 25 (4) AIA specifies further requirements for the exchange of information along the AI value chain, specifically between providers of a high-risk AI system and third parties that supply an AI system, tools, services, components, or processes that are used or integrated in a high-risk AI system. In these cases, the parties need to contractually agree on the necessary information, capabilities, technical access and other assistance to enable the provider of the high-risk AI system to fully comply with the AIA requirements. These obligations could apply to the relationship between deployers and providers of high-risk AI system, if the deployer itself becomes a provider of the high-risk system under the conditions stipulated by Art. 25 (1) AIA. For example, this may be the case when the deployer puts its name or trademark on the high-risk AI system without further contractual arrangements stating that the obligations are otherwise allocated.

[60] Panigutti, C., Hamon, R., Hupont, I., Fernandez Llorca, D., Fano Yela, D., Junklewitz, H., ... & Gomez, E. (2023). The role of explainable AI in the context of the AI Act. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1139-1150).

[61] Art 13 (1), Art. 13 (2) and Recital 72 AIA.

[62] See also Nannini, L. (2024). Habemus a Right to an Explanation: So What? – A Framework on Transparency-Explainability Functionality and Tensions in the EU AI Act. In *Proceedings of the Seventh AAAI/ACM Conference on AI, Ethics, and Society* pp. 1023-1035).

Information about the Characteristics, Capabilities and Limitations of Performance of a High-Risk AI System

Art. 13 (3)(b) AIA requires that the instructions for use of a high-risk AI system contain a list of information supposed to facilitate the deployers' understanding of the characteristics, capabilities and limitations of performance of the system. This includes information about the technical capabilities and characteristics of the high-risk AI system that are relevant to explain its output (where applicable).[63]

For trained machine learning models, "model cards" have been developed as a framework to document their intended use cases and performance characteristics.[64] These model cards have gained popularity in practice and are now supported by popular model hosting platforms and repositories as well as cloud service providers.[65] By integrating metadata into model cards programmatic discovery, evaluation and comparison of models can be facilitated. A model card should describe the employed model, its intended uses and potential limitations (including biases and ethical considerations), training parameters and information on testing and experimentation, the datasets used for training, and the model's evaluation results. As such, complete model cards can address transparency requirements specified by Art. 13 (3)(b) on the intended purpose of the AI system (point i), its level of accuracy and non-functional characteristics (point ii), its technical capabilities and characteristics (point iv), specifications for the input data and relevant information about training and test data (point vi), and possibly to some degree information to enable deployers to interpret its outputs (point vii). Therefore, effective and efficient operationalisation of transparency requirements could build on the existing concept and implementations of model cards.[66]

While conventional model cards typically focus exclusively on the AI model, achieving transparency on the system level requires incorporating additional information about how the AI model is integrated into the broader AI system, and how additional data processing at the system level contributes to the AI system's outputs. Hence, "system cards" may extend conventional model cards and include the aforementioned information in order to inform deployers about the characteristics, capabilities and limitations at the system layer. In this sense, "system cards" for a specific high-risk AI system are expected to be more focused and tailored to the corresponding intended use case than for example model cards for general-purpose AI models.

Although model cards may contain information about the potential risks associated with an AI system as well as the circumstances and factors that could contribute to these risks, these aspects are usually not the main focus of model cards in practice. Hence, to implement the transparency requirements of the AIA regarding the documentation of potential risks to the health, safety or fundamental rights (Art. 13 (3)(b)(iii) AIA), known or foreseeable circumstances that may negatively impact performance or

[63] Art. 13 (3)(b)(iv) AIA

[64] Art. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 220-229).

[65] See, for example, https://huggingface.co/docs/hub/en/model-cards, https://modelcards.withgoogle.com/about, https://docs.aws.amazon.com/sagemaker/latest/dg/model-cards.html

[66] Note that conventional model cards may focus exclusively on the AI model. Model cards or "system cards" describing an AI system may include additional information about the integration of the AI model into the broader AI system, and how additional data processing at the system level contributes to the AI system's outputs, should be provided.

non-functional characteristics (point ii), and the performance regarding specific persons or groups of persons on which the system is intended to be used (point v), additional information will regularly need to be provided to the deployer. This information could be generated as summaries from the more detailed information on the same issues as required by point 3 of Annex IV AIA on the technical documentation referred to in Art. 11 (1) AIA. Notably, the information should educate deployers about the intended and precluded uses of the AI system and to use the AI system correctly and as appropriate in light of the potential limitations and risks.[67] In practice, this information may also be part of the contract concluded between the provider and the deployer of a high-risk AI system. To promote accessibility and elucidate appropriate use cases and limitations more tangibly, illustrative examples should be used to support these summaries where feasible and appropriate.[68] To facilitate the consistent and accessible provision of information on risks associated with AI systems, recent efforts to augment the concept of model cards through risk documentation present a promising starting point for implementation.[69] This approach may also support contractual agreements between providers and deployers on the potential use of the high-risk AI system by providing a reference template.

In many areas of applications, AI systems will be regularly updated, and new versions will be released. To support the implementation of an effective transparency mechanism and to facilitate deployers' access to relevant information, the instructions for use for new releases of an AI system should contain information on the changes that have occurred in comparison to the last major version of the AI system. Such changelogs are widely used to support deployers in identifying new features or fixed security issues of newly released versions of AI systems or software more generally. This approach could be extended to cover changes in the identified risks and limitations of the AI system, thus supporting the transparency of issues and mitigation measures identified through continuous post-market monitoring.[70]

Performance Metrics and Non-Functional Characteristics of AI Systems

Reporting of performance metrics can help to characterise the capabilities and limitations of an AI system based on measurable criteria. Measurement of non-functional characteristics such as robustness or cybersecurity can further inform deployers about possible sources of risks. In this vein, Art. 13 (3)(b)(ii) emphasises that instructions for use should include information on the level of accuracy, including its metrics, robustness and cybersecurity against which the high-risk AI system has been tested and validated and which can be expected from the system.

While there is a myriad of performance metrics for AI systems and their models, specifically for accuracy as a primary performance criterion, the selection of these metrics and their appropriateness is often dependent on the employed technical approach and AI model as well as the specific context of the intended use of the AI system. Regarding the role of metrics to support the assessment of trustworthiness,

[67] Recital 72 AIA.

[68] Recital 72 AIA.

[69] See the proposal of "AI cards" building on the concept of model cards: Golpayegani, D., Hupont, I., Panigutti, C., Pandit, H. J., Schade, S., O'Sullivan, D., & Lewis, D. (2024). AI cards: Towards an applied framework for machine-readable ai and risk documentation inspired by the EU AI Act. In *Annual Privacy Forum* (pp. 48-72). Cham: Springer Nature Switzerland.

[70] In this vein, the proposed approach could support or possibly provide a template for contractual agreements on the communication of changes in the identified risks or limitations of the model between the provider and the deployer of a high-risk AI system.

it has been acknowledged that “the current lack of consensus on robust and verifiable measurement methods for risk and trustworthiness, and applicability to different AI use cases” presents an AI risk measurement challenge.[71] Hence, operationalisation of the requirements set out in Art. 15 AIA and referenced by the transparency requirements in Art. 13 (3)(b)(ii) AIA do not only raise open questions about the necessary and appropriate thresholds regarding these metrics, as further discussed in Section 4, but also to what extent the selection of metrics could be standardised to facilitate effective transparency and its efficient provision as part of the instructions for use.

Different areas of application can thereby present fundamentally different challenges to the establishment of adequate metrics and their measurement (such as for accuracy). For example, in medical contexts, AI systems can often be applied to and validated against test cases that have a measurable and objective outcome, such as if the choice for a particular drug led to a healthier outcome or whether the pattern classified in an image correctly identified a symptom of a disease.[72] Even in such cases, there remain measurement challenges due to natural variability and noisy signals, but those can be mitigated by well-known statistical methods, at least to some extent.[73] However, in other high-risk contexts, even more fundamental challenges arise to measuring metrics such as for accuracy, as there exists no clear benchmark for the “ground truth” of what is an accurate output or decision (either by an AI system or a human). For example, in the workplace context, AI systems are already used to automate or support humans in hiring employees for organisations, for example, by selecting suitable candidates based on the analysis of written applications and curriculum vitae or the evaluation of pre-recorded presentation videos. Whether a candidate is indeed the accurate choice is very difficult to assess, as it is already unclear on what basis and according to which perspective one would determine a good or bad choice in this context.

This illustrates only one, although fundamental challenge to establishing uniform performance criteria and metrics on the universal level of the AIA as a horizontal regulation. The significant heterogeneity in contexts and use cases across risk domains, combined with the diverse technical approaches underlying AI systems, suggests that guidelines and standards for operationalising performance measurement should be developed at the level of specific high-risk domains. The classification of high-risk systems according to Art. 6 (1) and Art. 6 (2) may provide a framework for the high-level delineation of these risk domains, although even more granular delineation may be necessary in broad domains. This would facilitate the selection of a consistent set of performance criteria and metrics targeted to the use cases of AI systems in the corresponding risk domain. Such an approach would also be conducive to the derivation of more tangible proportionate thresholds that should be met by the performance of high-risk AI systems for specific use cases and contexts, as verified by the corresponding performance metrics. Inherently, this more domain-specific approach raises challenges regarding additional complexity and dependencies,

[71] NIST (2023), p. 6.

[72] See, for example, US Food and Drug Administration (2018). FDA permits marketing of artificial intelligence-based device to detect certain diabetes-related eye problems. Available at https://www.fda.gov/news-events/press-announcements/fda-permits-marketing-artificial-intelligence-based-device-detect-certain-diabetes-related-eye; McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., ... & Shetty, S. (2020). International evaluation of an AI system for breast cancer screening. *Nature*, *577*(7788), 89-94.

[73] See, for example, Liu, X., Cruz Rivera, S., Moher, D., Calvert, M. J., & Denniston, A. K. (2020). Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nature Medicine*, *26*(9), 1364-1374.

avoiding sectoral fragmentation and ensuring horizontal consistency. However, as outlined above, this step seems necessary to establish actionable common practices for implementation beyond the abstract principles and rules that can be derived on the horizontal level.

Testing Procedures and Data Sets Used for Training and Validation

Art. 13 (3)(b)(v) AIA requires that, when appropriate, the instructions for use include specifications for the input data, or any other relevant information in terms of the training, validation and testing data sets used, taking into account the intended purpose of the high-risk AI system. Transparency about the testing procedures and the test samples can serve as additional information for deployers to evaluate and contextualise the performance metrics and non-functional characteristics reported by the provider. This is important because measurements of performance metrics and non-functional characteristics depend critically on the performed testing procedures and employed datasets. In general, the scope of testing procedures should consider the criteria established for risk management systems of high-risk AI systems in Art. 9 (6), (7) and (8) AIA.[74] Being able to assess the scope of testing and the suitability and generalisability of the data helps deployers assessing the trustworthiness and robustness of the supposed performance of the high-risk AI system. Transparency about the testing procedures and datasets establishes further information on the scope of evaluation, that is, the breadth of use cases and the variability of different influencing conditions that the AI system was evaluated against. Consequently, this allows deployers to assess the acceptable boundaries of the use of the AI system and enable better judgements about the adequacy of using the systems in the deployer's domain and in the intended context.

Furthermore, testing and validation of high-risk AI systems are important elements in identifying risks that specific subgroups of people may be exposed to. Therefore, documentation of the testing procedures and employed data sets should contain information to which extent those covered different groups of persons. This information helps deployers assessing the performance of the AI system regarding the specific target audience on which the system is intended to be used, as required by Art. 13 (3)(b)(v). Furthermore, it can promote deployers' awareness of group-specific risks and potential limitations of the AI system (such as potential bias) that need particular attention or additional precautionary measures.

As for the reporting of performance measures, the level of detail in which testing procedures and employed datasets are described will be critical for effective transparency. In particular, the level of detail will determine to what extent this information can inform deployers' understanding of a system's capabilities and limitations or improve the interpretability of its model outputs. Determining the appropriate level of detail involves the same general trade-offs as discussed before on p. 21 and should account for the principle of proportionality. To establish more specific and readily implementable criteria for the transparency of testing procedures and training data, guidelines or standards may be developed

[74] According to Art. 9 (6) AIA, this includes testing for identifying the most appropriate and targeted risk management measures as well as ensuring that the high-risk AI system performs consistently for the intended purpose and that it complies with the requirements set out in Section 2 of Chapter III AIA. Additionally, testing procedures may include testing in real-world conditions (Art. 9 (7) and Art. 60 AIA). Art 9 (8) AIA specifies the timing of testing and requires that testing should be carried out against appropriate, prior defined metrics.

for different risk domains, similar to those proposed for the transparency of performance measurement. The development of domain-specific criteria may then consider the extent to which general heuristics or rules of thumb for adequate testing are applicable and appropriate.[75] In addition, this approach allows for more specific guidelines on what testing methods would be deemed necessary for specific domains or use cases (such as red-teaming for identifying security vulnerabilities).[76]

These efforts could also aim to establish standardised testing datasets and procedures for specific use cases against which models can be evaluated.[77] Combined with common guidelines for reporting this information would allow deployers to access directly comparable and transparent information on testing and performance of high-risk AI systems in a specific context. Involving the AI Office (and possibly national competent authorities and notified bodies) in the development of such standardised testing procedures and test cases could further reduce uncertainty for providers of high-risk AI systems. To avoid overfitting to standardised datasets and excessive tailoring of the design and analysis of AI systems to specific test data as well as to account for shifting conditions and concept drifts, standardised data sets should not be static but change over time and be updated accordingly. In general, standardisation efforts must consider the dynamic development of AI technologies and hence should establish mechanisms for providers of high-risk AI systems to react timely to these developments without being unduly constrained by possibly lengthy procedures to update a standard's specifications.

3.3 AIA Requirements on the Interpretability of AI System Outputs

As part of the information about the characteristics, capabilities and limitations of performance of a high-risk AI system, Art. 13 AIA also requires that the instructions for use, where applicable, enable deployers to interpret the output of the high-risk AI system and use it appropriately.[78] In addition, Art. 13 (3)(d) AIA imposes transparency requirements regarding the human oversight measures set out in Art. 14 AIA. This includes information on "the technical measures put in place to facilitate the interpretation of the outputs of the high-risk systems by the deployers".[79] By referencing the interpretability of outputs, these requirements can be viewed to align with the definition of interpretability under the NIST AI RMF as referring to the "meaning of AI systems' output in the context of their designed functional purposes."[80]

Transparency on the elements, functions, characteristics, capabilities and limitations of an AI system can establish some degree of interpretability in the narrow sense, as such information can help humans to

[75] For example, it has been suggested that there should be a "minimum order of magnitude (10X) more training data than degrees of freedom of the AI model" as well as "a minimum of 1000 training samples per decision class"; see Petkovic, D. (2023). It is not "Accuracy vs. Explainability"—we need both for trustworthy AI systems. *IEEE Transactions on Technology and Society*, *4*(1), 46-53, p. 48.
[76] NIST (n.d.). Red Team. Available at https://csrc.nist.gov/glossary/term/red_team
[77] Publishing standardised datasets can be prone to overfitting or deliberate efforts to train and tune the performance of AI systems specifically on these datasets. However, there are approaches to mitigate such adverse effects, for example, by providing automated testbeds without disclosing the entire dataset, by randomising data inputs or by dynamically updating the dataset over time.
[78] Art. 13 (3)(b)(vii) AIA.
[79] Art. 13 (3)(d) AIA.
[80] NIST (2023), p. 16.

infer why the model made some specific output for a given instance. In addition, risks to interpretability can be addressed by additionally communicating a description of why an AI system generated a particular output or came to a particular decision.[81] The academic literature on Explainable AI suggests that techniques to improve the interpretability of outputs of AI systems could contribute to risk mitigation by promoting a range of benefits. In particular, the goals of making the outputs of AI systems more interpretable through such techniques include:[82]

1. *Trustworthiness*, defined as the confidence of whether a model will act as intended when facing a given problem.
2. *Informativeness,* related to extracting information about the inner relations of a model to provide a simpler understanding of what the model internally does.
3. *Transferability*, with specific regard to elucidating the boundaries of a model; in particular, this should help avoid cases in which the lack of a proper understanding of the model might drive the user toward incorrect assumptions and fatal consequences.
4. *Fairness*, for example, by facilitating the identification of bias affecting the output of an AI system stemming from the data the model was exposed to.
5. *Causality*, that is, facilitating the identification and evaluation of whether an AI system output is derived from a causal relationship as opposed to mere correlations.
6. *Confidence*, as a generalisation of whether an AI system is reliable in the sense that its outputs are stable and robust in the context of small variations of its inputs.

As mentioned above, there are two general approaches to achieving interpretability in the narrow sense of AI system outputs: intrinsically interpretable AI models and post-hoc explanations for black-box AI models.

Intrinsically Interpretable Models

Some types of AI models are intrinsically interpretable (also referred to as “transparent models”) meaning that they allow humans to directly discern the relationships or rules according to which inputs are transformed into outputs by the model without the need for further explanations or observations of input-output instances. To allow for such intrinsic interpretability these models often exhibit a simpler structure and usually do not include complex interactions of nonlinear transformations of inputs to generate outputs. For example, the model weights of a logistic regression model or the structure of a decision tree and its visual representation have natural interpretations that make the process by which an AI system building on these models transforms inputs into outputs understandable for humans. Most notably, these models therefore allow a human, ideally for any possible input instance, to anticipate how the output of a model will depend on the features (that is, the attributes) of the input instance and how the output would change in reaction to changes in the features of the input instance.

[81] NIST (2023), p. 17.

[82] The list consists of selected issues from the goals identified by Arrieta et al. (2020) in their survey of the literature on explainable AI.

While the empirical measurement of interpretability remains an unresolved and significant challenge, different levels of intrinsic interpretability in machine learning models have been proposed from a conceptual perspective (in increasing order):[83] (i) algorithmic transparency refers to the ability of the user to understand the process followed by the model to produce any given output from its input data. From a more technical perspective, this requires that the model must be fully explorable by means of mathematical analysis and methods; (ii) decomposability (or intelligibility) further requires that every part of the model, including the inputs, parameters and calculations must be understandable by a human without the need for additional tools; (iii) simulatability denotes the ability of a model to be cognitively "simulated" by a human, thus allowing the human to think and reason about the model in its entirety in addition to understanding all of its components.

However, the benefits from improved interpretability regularly come at the cost of lower accuracy and performance of intrinsically interpretable models compared to the best-performing black-box AI models. This has led to the notion of a general accuracy-explainability trade-off.[84] While some proponents of interpretable models contest the existence of such a trade-off on a fundamental level,[85] and more sophisticated intrinsically interpretable models such as generalised additive models[86] have shown some promising signs to narrow the accuracy gap relative to leading black-box models, the success of deep neural networks and the widespread adoption of black-box models across use cases indicate that performance advantages of the latter generally persist. Furthermore, as the performance gain from black-box models comes from their ability to capture the complex relationships and patterns of the underlying phenomena, capturing the same phenomena by simpler, human-interpretable models necessarily requires a reduction of model complexity, which will require some compromise of general model accuracy.

The performance that intrinsically interpretable models can achieve will also depend critically on the type of input data. Whereas specialised interpretable models may achieve relatively high accuracy on tabular data, processing of video or text data presents significantly greater challenges for simpler models to match the accuracy of more complex models, such as generative pre-trained transformers used in state-of-the-art LLMs. Interestingly, the recent success of LLMs has inspired new research into whether such black-box models can augment intrinsically interpretable models to improve the performance of simpler models, with promising early results.[87]

[83] Arrieta et al. (2020).

[84] DARPA (2016). Broad Agency Announcement**.** Explainable Artificial Intelligence (XAI). DARPA-BAA-16-53. Available at https://www.darpa.mil/attachments/DARPA-BAA-16-53.pdf.

[85] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, *1*(5), 206-215.

[86] Hastie, T., & Tibshirani, R. (1986). Generalized Additive Models. *Statistical Science*, *1*(3), 297-310; Lou, Y., Caruana, R., & Gehrke, J. (2012). Intelligible models for classification and regression. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 150-158).

[87] Singh, C., Askari, A., Caruana, R., & Gao, J. (2023). Augmenting interpretable models with large language models during training. *Nature Communications*, *14*(1), 7913.; Singh, C., Inala, J. P., Galley, M., Caruana, R., & Gao, J. (2024). Rethinking interpretability in the era of large language models. Available at https://doi.org/10.48550/arXiv.2402.01761.

Post-hoc Explanations for Black-Box AI Models

Given the success of deep learning over the last decade, many contemporary AI systems are built on top of black-box models that are not intrinsically interpretable, as humans regularly cannot anticipate the output for a given input from the parameters of the trained model and its structure. To improve the interpretability of these AI systems and their output, research on Explainable AI has developed a range of post-hoc explanation techniques that aim to improve humans' understanding of how and why these systems produce a certain output. For instance, explanations based on feature importance aim to identify which features played a significant role in determining the output of the AI system and what effect the value of a feature has on the output (for instance, whether the feature value makes it more or less likely that the instance is categorised into a certain output class or whether a feature value leads to a lower or higher numerical prediction). These approaches often rely on surrogate simple models to approximate and explain the behaviour of a black-box AI model. Popular approaches include LIME and SHAP, which are model-agnostic and can thus, in principle, be applied to any black-box model.[88] Explanations may be local, that is, an explanation for each individual output of an AI model, or global, that is, for the overall (average) behaviour of the model. Often, global explanations are derived by aggregating over a large number of representative local explanations.

While post-hoc explanations have great appeal because they promise to combine the high predictive accuracy of black-box AI models with the interpretability of their outputs, there are several limitations to their universal applicability. When post-hoc explanation techniques implement simpler surrogate models to approximate more complex black-box AI models, they may sometimes produce explanations that are not consistent with the output and the actual processing of the original model.[89] Therefore, these techniques can be prone to producing inaccurate explanations of the model output, a limitation commonly referred to in the academic literature as a lack of faithfulness.[90] Providing inaccurate, but plausible explanations to humans is particularly problematic, as it has been shown that humans adapt their decision-making to explanations, not only for the specific input instance for which the explanation was presented, but also for subsequent instances.[91] Even faithful explanations may lead to adverse side effects, as humans' cognitive processing and acceptance of explanations can be prone to behavioural biases. For example, it has been found that humans are more likely to follow explanations of an AI system that confirm their prior beliefs, while paying less attention to contradictory explanations.[92]

[88] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). "Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135-1144). Lundberg, S. M., & Lee, S. I. (2017). A Unified Approach to Interpreting Model Predictions. In *Proceedings of the 31st Conference on Neural Information Processing Systems* (pp. 4765-4774).

[89] Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, *3*(11), e745-e750.

[90] Jacovi, A., & Goldberg, Y. (2020). Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

[91] Bauer, K., von Zahn, M., & Hinz, O. (2023). Expl(AI)ned: The impact of explainable artificial intelligence on users' information processing. *Information Systems Research*, *34*(4), 1582-1602; Morrison, K., Spitzer, P., Turri, V., Feng, M., Kühl, N., & Perer, A. (2024). The Impact of Imperfect XAI on Human-AI Decision-Making. *Proceedings of the ACM on Human-Computer Interaction*, *8*(CSCW1), 1-39.

[92] Bauer, von Zahn & Hinz (2023).

A further challenge for post-hoc explanation techniques is the reliability of explanations in terms of their stability across different input instances and their robustness. For example, a well-known limitation of LIME is that this method may generate different or even contradictory explanations for the same input instance, and that the provided feature importance values depend critically on the precise implementation of the method.[93] Moreover, post-hoc explanations can be vulnerable to adversarial attacks, posing the risk that malicious actors or developers of an AI system could manipulate explanations.[94] The application of post-hoc explanation techniques can also be complicated by their computational requirements. Especially regarding global explanations, some techniques face high computational costs, especially if the input data is of high dimensionality and thus exhibits a very large number of features.[95]

The potential limitations of current post-hoc explanation techniques have sparked vivid research on the evaluation and validation of the adequacy of existing techniques as well as the development of new approaches and methods that aim to address these limitations.[96] From a practical perspective, this strand of research is promising, as it elucidates the factors, contexts and risk domains where specific post-hoc explanation techniques can effectively improve interpretability, as well as where existing approaches fall short. Overall, this suggests that post-hoc techniques hold potential for improving the interpretability of black-box models. However, their application currently requires careful vetting and validation to ensure their adequacy for the intended use case and its specific conditions as well as to avoid adverse effects.

Implementing AIA Requirements on the Interpretability of AI System Outputs

The previous discussion of technical approaches indicates that there is currently no universal method for achieving interpretability of AI systems' outputs in the narrow sense, whether through intrinsically interpretable models or post-hoc explanation techniques. While intrinsically interpretable models generally suffer from lower accuracy and performance relative to black-box AI models, post-hoc explanation techniques may not always be accurate, reliable or computationally feasible, among other challenges.[97]

Among scholars, this has led to a debate with diverse views on how to reconcile the tension between the goal of achieving interpretability of AI systems' outputs and the limitations of technical methods in

[93] Molnar, C. (2024). Local Surrogate (LIME). In *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable;* Alvarez-Melis, D., & Jaakkola, T. S. (2018). On the robustness of interpretability methods. *ICML Workshop on Human Interpretability in Machine Learning (WHI 2018)*, Stockholm, Sweden.

[94] Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2020). Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 180-186).

[95] For example, SHAP explanations are generally considered more robust than LIME, but are significantly more computationally expensive to calculate. See Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., ... & Ranjan, R. (2023). Explainable AI (XAI): Core ideas, techniques, and solutions. *ACM Computing Surveys*, *55*(9), 1-33.

[96] See, for example, Kim, B. R., Srinivasan, K., Kong, S. H., Kim, J. H., Shin, C. S., & Ram, S. (2023). ROLEX: A Novel Method for Interpretable Machine Learning Using Robust Local Explanations. *MIS Quarterly*, *47*(3); Heinrich, B., Krapf, T., & Miethaner, P. (2024). EXPLORE: A Novel Method for Local Explanations. In *Proceedings of the Forty-Fifth International Conference on Information Systems (ICIS 2024)*.

[97] Additional challenges for the interpretability of system AI outputs can arise from the model's integration into the system and additional data processing at the system level. In cases where the AI system provider and the AI model provider are separate entities, this adds even more complexity to ensuring interpretability.

establishing such interpretability. Many researchers generally acknowledge the need for interpretability especially for high-stakes AI applications.[98] However, some advocate for the use of black-box models even when they lack interpretability, as they argue that the benefits from more accurate outputs can be more important than the ability to explain them.[99] In addition, it is highlighted that human decision-making as an alternative to black-box AI models can often be similarly opaque, with its outputs being equally challenging for others to interpret.[100] In contrast, others have advocated for the exclusive use of interpretable models in high-stakes AI applications and to abandon black-box models altogether.[101] Then again, some scholars have suggested that the matter is even more complex, as some intrinsically interpretable models may perform worse than black-box models in facilitating humans' understanding of the model output when humans' ability to anticipate outputs is measured empirically.[102]

The latter finding points to a more general challenge in measuring and validating interpretability in the narrow sense. While there exist theoretical concepts and an intuitive understanding of what interpretability of AI system outputs refers to, quantitative and objective measurement of interpretability is difficult. As a consequence, it is equally difficult to define a threshold that would indicate sufficient interpretability. This is further complicated by the well-known fact that interpretability depends crucially on human understandability and thus on a specific audience and its subjective understandability shaped by its cognitive skills and pursued goals.[103] In addition, interpretability is typically a means to ensure other goals and support different purposes of risk mitigation that can vary across risk domains and use cases. In this context, it is important to note that improving interpretability can involve trade-offs with other non-functional goals for high-risk AI systems, such as fairness, which are supposed to be promoted through interpretability in the context of the AIA.[104]

Because of the limitations of interpretability in the context of complex machine learning models, quantifying and communicating the uncertainty of an AI system regarding a specific output may represent a reasonable and, in some cases, more effective measure to mitigate the risks of concern or to support human oversight.[105] Hence, methods for uncertainty quantification should be considered and evaluated

[98] Petkovic, D. (2023). It is not "Accuracy vs. Explainability"—we need both for trustworthy AI systems. *IEEE Transactions on Technology and Society*, *4*(1), 46-53.

[99] London, A. J. (2019). Artificial intelligence and black-box medical decisions: accuracy versus explainability. *The Hastings Center Report*, *49*(1), 15-21; Holm, E. A. (2019). In defense of the black box. *Science*, *364*(6435), 26-27.

[100] In this context, it is important to note that although human decision-making may sometimes be similarly opaque, there are established mechanisms to hold humans accountable. However, these mechanisms are not directly applicable to AI models or systems.

[101] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, *1*(5), 206-215.

[102] Bell, A., Solano-Kamaiko, I., Nov, O., & Stoyanovich, J. (2022). It's just not that simple: An empirical study of the accuracy-explainability trade-off in machine learning for public policy. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 248-266).

[103] See Arrieta et al. (2020) and Bauer, von Zahn & Hinz (2023).

[104] See, for example, Jabbari, S., Ou, H. C., Lakkaraju, H., & Tambe, M. (2020). An empirical study of the trade-offs between interpretability and fairness. In *ICML Workshop on Human Interpretability in Machine Learning, International Conference on Machine Learning (ICML)*.

[105] Hüllermeier, E., & Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, *110*(3), 457-506; Bobek, S., & Nalepa, G. J. (2021). Introducing uncertainty into explainable AI methods. In *International Conference on Computational Science* (pp. 444-457). Springer International Publishing.

as possible alternatives when interpretability is difficult to ensure or cannot effectively mitigate the targeted risk.

As there is no general solution to the problem of resolving the trade-off between interpretability and accuracy of AI systems as well as tensions with other non-functional goals, these can be navigated and balanced most effectively by considering the specific use cases of a high-risk AI system and its application context. This conclusion has also been emphasised by other scholars who argue, for instance, that the performance-explainability trade-off is “best approached in a nuanced way that incorporates resource availability, domain characteristics, and considerations of risk.”[106] Moreover, this perspective aligns with recent calls to evaluate the suitability of methods for improving interpretability of AI system outputs in the narrow sense based on their ability to satisfy the desiderata of affected stakeholders. [107]

In the context of the AIA, this would indicate the need for a more specific evaluation of a use case or high-risk AI system and the associated risks to decide whether methods aimed at promoting interpretability in the narrow sense are feasible, necessary, appropriate and proportionate. This aligns with recommendations in other risk frameworks. For example, the NIST AI RMF states that “when consequences are severe, such as when life and liberty are at stake, AI developers and deployers should consider proportionally and proactively adjusting their transparency and accountability practices.”[108] This perspective does not neglect the general requirement of the AIA regarding the interpretability of outputs of high-risk AI systems. Instead, it suggests that transparency about the AI system, as referred to in Section 3.2 can enable sufficient interpretability of outputs of an AI system, unless risks associated with this system can only be mitigated by enabling methods aimed at improving interpretability in the narrow sense. This perspective is supported by Panigutti et al., who analyse the role of explainable AI in the context of the initial AIA draft proposed by the Commission.[109] Based on their interdisciplinary analysis of the AIA proposal, they argue that the AI Act does not mandate a requirement for explainable AI methods aimed at establishing interpretability in the narrow sense and that, instead, transparency, documentation and human oversight measures represent the primary measures to achieve the goals of the AIA.

To promote effective and efficient implementation, guidelines or standards may specify the conditions or use cases that would require the mandatory use of intrinsically interpretable models or post-hoc explanation techniques in a specific risk domain. Such requirements on interpretability in the narrow sense should be preceded by an analysis of the feasibility and adequacy of interpretability methods, the specific trade-offs with other performance and non-performance goals, as well as their effectiveness in mitigating the specific risks.

[106] Crook, B., Schlüter, M., & Speith, T. (2023). Revisiting the performance-explainability trade-off in explainable artificial intelligence (XAI). In *2023 IEEE 31st International Requirements Engineering Conference Workshops (REW)* (pp. 316-324). IEEE.

[107] Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., ... & Baum, K. (2021). What do we want from Explainable Artificial Intelligence (XAI)? – A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence*, *296*, 103473.

[108] NIST, 2023, p. 16.

[109] Panigutti, C., Hamon, R., Hupont, I., Fernandez Llorca, D., Fano Yela, D., Junklewitz, H., ... & Gomez, E. (2023). The role of explainable AI in the context of the AI Act. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1139-1150); Proposal 2021/0106 of 21 April 2021 for a regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52021PC0206

3.4 Approaches Towards Effective and Efficient Implementation

Based on the preceding analysis, the following recommendations are derived for approaches towards effective and efficient implementation of the AIA's requirement for transparency of high-risk AI systems and the interpretability of their outputs. While these recommendations focus on a specific AIA requirement, their underlying principles may inform also the implementation of other AIA requirements. As Section 2 of Chapter III of the AIA covers a diverse set of risk domains, but also AI systems associated with different risks and levels of potential harm, a general principle applied here is that implementation requirements should aim to keep entry barriers for introducing new AI systems low, unless known or foreseeable risks for the specific domain or use case warrant stricter requirements. Such stricter requirements may involve the provision of more detailed information, the mandatory use of specific metrics or procedures, or the application of methods to enhance interpretability in the narrow sense.

To reduce uncertainty about whether and when stricter implementation requirements are deemed necessary and expected to be effective, codes of conduct, guidelines, and standards should be established on the level of individual risk domains. This would facilitate the derivation of implementation requirements tailored to more specific contexts and use cases. At the same time, cross-cutting principles and general rules are necessary to ensure consistency between risk domains, avoid sectoral fragmentation and minimise compliance cost.[110]

To further ensure the effective mitigation of risks associated with high-risk AI systems, the importance of continuous compliance and post-market monitoring is emphasised. As ex-ante anticipation and testing of all possible risks is costly and often unfeasible, incorporating feedback from deployers at the provider level and addressing emerging risks collaboratively across the AI value chain is particularly relevant. Therefore, effective and efficient implementation should also include the establishment of adequate procedures and institutions that can facilitate the exchange of relevant information and promote coordination among operators of AI systems. These procedures and institutions should build and support bilateral contractual agreements along the AI value chain as required by Art 25 (4) AIA.

Implementation of AIA Transparency Requirements

The analysis of the AIA requirements on transparency of high-risk AI systems and the interpretability of their outputs suggests that the instructions for use should cover three main categories of information:

1. **Information on characteristics, capabilities and limitations of performance of high-risk AI systems.**
   To provide this information the instructions for use of high-risk AI systems may build on the concept of model cards, which are already widely used to describe the AI model employed in an AI system, its intended use and potential limitations, training parameters and information on

[110] See, for example, the proposed principle on appropriate accuracy in Section 4.2, which could represent a general rule that can be further contextualised in individual risk domains.

testing and experimentation, the datasets used for training, and the model's evaluation results. This approach may complement and support contractual agreements between providers and deployers on the potential use of the high-risk AI system. A key question for implementation concerns the necessary level of detail that needs to be provided regarding this information, specifically for information on performance measurement, testing procedures and the employed training data. While there are general principles to determine the appropriate level of detail, effective and efficient implementation could be promoted by establishing common practices through guidelines or standards at the level of risk domains. This also presents opportunities to specify common performance metrics and testing procedures or develop standardised test datasets for more specific use cases, accounting for the conditions of a particular risk domain.[111]

2. **Documentation of potential risks and known or foreseeable circumstances that negatively impact functional or non-functional characteristics of the AI system.**
   This additional information is necessary to inform deployers about the potential risks of an AI system and negative influencing factors to allow for effective risk mitigation at the downstream level. The provision of this information could be facilitated by augmenting model cards to support the consistent documentation of potential risks and relevant influencing factors. As ex-ante anticipation of all possible risks is difficult, if not impossible, implementation of the AIA requirements should support effective post-market monitoring and information exchange as well as coordination mechanisms between deployers and providers of high-risk AI systems building on contractual agreements along the AI value chain as set out in Art. 25 (4) AIA.

3. **Information and measures to enable interpretability of AI system outputs**.
   In general, the transparency information outlined in the two preceding categories can provide a sufficient level of interpretability for the outputs of an AI system, as required by the AIA, as such information can help humans infer why the model made some specific output for a given instance. In specific cases, additional measures to enhance interpretability in the narrow sense may be warranted to address severe risks associated with a particular high-risk AI system or use case. Guidelines or standards for specific risk domains may identify such use cases and also establish whether intrinsically interpretable models or post-hoc explanation techniques should be considered suitable measures to address the risks in these cases. This requires an analysis of the feasibility and adequacy of these methods in the given context, the trade-offs with other performance and non-performance goals, as well as their effectiveness in mitigating the specific risks.

[111] In turn, this poses the question, what level of granularity can still offer benefits over costs from excessive fragmentation. This question will need to be addressed in line with the institutional framework that is emerging for the implementation of the AIA. See Larouche (2025).

Establishing Common Practices Through Guidelines and Standardisation for Specific Risk Domains

The AIA envisions standards, codes of conduct and guidelines as tools for implementation.[112] Regarding implementation, standardisation should play a key role to provide technical solutions to providers to ensure compliance.[113] As a subsidiary measure, Art. 41 AIA outlines common specifications for the requirements set out in Section 2 of Chapter III AIA, when no appropriate harmonised standard is available or established according to the Criteria of Art. 41 (1)(a) AIA.

In the context of the transparency requirements of the AIA, common practices can promote effectiveness and efficiency, as deployers can more easily assess and evaluate provided information as well as compare different AI systems. At the same time, this can help providers of AI systems by reducing uncertainty and costs for compliance by making it possible to resort to common practices. However, the broad scope of the AIA requirements makes it difficult to achieve such common practices in a tangible and effective manner, given the various specificities of the diverse use cases and associated technical approaches underlying different high-risk AI systems. Hence, standards or guidelines could be more effective, when developed for specific risk domains of high-risk AI systems to promote common practices for implementation. The classification of high-risk systems as specified in Art. 6 (1) and Art. 6 (2) AIA may provide a general delineation of these different application domains, although more granular approaches may be necessary in broad domains.[114] Notwithstanding, there may be common principles and criteria that cut across risk domains. However, even standards and guidelines for these cross-cutting principles may benefit from a bottom-up approach building on domain-specific analyses and best practices.

As mentioned above, common practices through guidelines or standards at the level of risk domains can help identify use cases where interpretability of AI system outputs in the narrow sense is deemed essential. They may then also establish whether intrinsically interpretable models or post-hoc explanation techniques should be considered suitable and effective measures in these cases. Furthermore, establishing common practices for specific risk domains can promote the effective implementation of transparency requirements by specifying common performance metrics and testing procedures, or developing standardised test datasets with regard to specific use cases and conditions in a particular risk domain. To this end, domain-specific standards could build on emerging general standards on the transparency of AI systems, such as ISO/IEC 12792, a draft standard for a transparency taxonomy.[115]

[112] Art. 40, 95, and 96 AIA respectively.

[113] Recital 121.

[114] At the same time, this needs to be balanced with the risk of an overly fragmented approach when identifying too narrow risk domains.

[115] Soler Garrido, J., Fano Yela, D., Panigutti, C., Junklewitz, H., Hamon, R., Evas, T., ... & Scalzo, S. (2023). Analysis of the preliminary AI standardisation work plan in support of the AI Act. JRC Technical Report. Available at https://publications.jrc.ec.europa.eu/repository/handle/JRC132833; ISO/IEC DIS 12792. Information technology — Artificial intelligence — Transparency taxonomy of AI systems. Available at https://www.iso.org/standard/84111.html .

Timely Transparency Through Post-Market Monitoring of Risks and Effective Feedback Mechanisms Across the AI Value Chain

As part of the transparency requirements, the AIA requires providers of high-risk AI systems to inform deployers about the known or foreseeable circumstances related to its use that may lead to risks to the health and safety or fundamental rights. In general, these risks should be identified through the provider's risk management system. As Art. 9 (2)(c) AIA mentions, this includes risks identified through post-market monitoring. Implementing effective post-market monitoring appears particularly important for achieving the goals of the AIA. This is because anticipating all risks before the deployment of an AI system is particularly challenging. First, new types of risks may emerge only after the deployment of the AI system.[116] Second, risks may change over time as data inputs or the environment of the deployed AI system evolve. Third, the enumeration of all possible risks through pre-deployment testing is costly and often unfeasible. Finally, risks may arise from the combination of third-party software, hardware, and data as well as interactions between AI systems, becoming observable only after deployment.

Therefore, the AIA acknowledges the value of post-market monitoring to evaluate the continuous compliance of AI systems with the requirements set out in Section 2 of Chapter III AIA and outlines specific obligations in Art. 72 AIA. It is further mentioned that the Commission shall adopt an implementing act laying down detailed provisions establishing a template for the post-market monitoring plan by February 2026.[117]

As deployers are the actors most likely to first observe risks and limitations of high-risk AI systems arising after their deployment, effective post-market monitoring should be supported by feedback mechanisms that allow and encourage deployers to report any relevant observations and incidents to the provider of a high-risk AI system.[118] In addition, the provider of a high-risk AI system may rely on information obtained from deployers through specific requests to identify and evaluate potential risks or to further test mitigation strategies.

Such feedback mechanisms should complement mandatory reporting obligations for serious incidents as established by the AIA, which need to be reported to the market surveillance authorities of the Member States where that incident occurred.[119] Here, a serious incident refers to particularly severe harm, infringement or disruption and ensuring that public authorities are informed about these incidents.[120] In contrast, the envisioned feedback mechanisms for post-market monitoring aim for a collaborative approach between providers and deployers to anticipate and mitigate emerging risks, consider limitations of performance and possible remedies that arise with respect to context-specific applications, or identify changes in the inputs and environments that could affect the performance of the AI system. The information gathered through such feedback mechanisms should then be dynamically incorporated into

[116] See also Weidinger, L., Rauh, M., Marchal, N., Manzini, A., Hendricks, L. A., Mateos-Garcia, J., ... & Isaac, W. (2023). Sociotechnical safety evaluation of generative AI systems. Available at https://doi.org/10.48550/arXiv.2310.11986.

[117] Art. 72 (3) AIA.

[118] In this vein, feedback mechanisms and institutions should support implementation of Art. 26 (5) AIA.

[119] Art. 73 AIA.

[120] Art. 3 (49) AIA.

the instructions for use of the concerned high-risk AI system and made accessible to all deployers to promote timely transparency.

4 Requirement for Appropriate Accuracy

Art. 15 (1) AIA requires that "[h]igh-risk AI systems shall be designed and developed in such a way that they achieve an appropriate level of accuracy, robustness and cybersecurity, and that they perform consistently in those respects throughout their lifecycle." Thus, the AIA emphasises that high-risk AI systems must exhibit appropriate performance in terms of accuracy and appropriate levels of non-functional criteria. The discussion here focuses on the requirement of appropriate accuracy.

As part of the risk management system, testing must ensure that high-risk AI systems comply with the requirement for appropriate accuracy.[121] Before a high-risk system is placed on the market or put into service, providers are required to provide detailed information about its capabilities and limitations in performance as part of the mandatory technical documentation.[122] This information should include "the degrees of accuracy for specific persons or groups of persons on which the system is intended to be used and the overall expected level of accuracy in relation to its intended purpose". In addition, the technical documentation should provide information on the validation and testing procedures as well as the metrics used to measure accuracy.[123]

Furthermore, as described in Section 3, the levels of accuracy and the relevant metrics of a high-risk AI system shall be communicated to the deployers of high-risk AI systems as part of the instructions for use issued by the provider.[124] The level of detail regarding the accuracy of a high-risk AI system may differ between the technical documentation and the instructions for use.

4.1 Standards for Benchmarks and Measurement Methodologies for the Accuracy of High-risk AI Systems

Art. 15 (2) AIA considers the technical aspects of measuring the appropriate level of accuracy and other relevant performance metrics. Specifically, the Commission, in cooperation with relevant stakeholders and standard-setting organisations, shall "encourage, as appropriate, the development of benchmarks and measurement technologies" to address these technical aspects.

Art. 40 (1) stipulates that high-risk AI systems conforming to harmonised standards, or parts thereof, are presumed to comply with the requirements set out in Section 2 of Chapter III AIA, including the obligation of ensuring appropriate accuracy. As a subsidiary measure, the Commission may adopt implementing acts to establish common specifications for these requirements in cases when no appropriate harmonised standard is available or established according to the Criteria of Art. 41 (1)(a) AIA.[125]

[121] Art. 6 (6) AIA.
[122] Art. 11 (1) and Annex IV (3) AIA.
[123] Annex IV (2)(g) AIA.
[124] Art 15 (3) and Art. 13 (2)(ii) AIA.
[125] Art. 41 AIA. See Larouche (2024) for a discussion of the potential tensions and uncertainties when the Commission will need to decide whether a harmonised standard is to be published pursuant to Art. 10 of Regulation 1025/2012.

Given the key role of standards, the Commission has engaged with European Standardisation Organisations since the proposal for the AI Regulation was initially presented.[126] In May 2023, the Commission adopted a formal standardisation request, which was accepted by the European Committee for Standardization and the European Committee for Electrotechnical Standardization (CEN-CENELEC).[127] The standardisation request includes "specifications for ensuring an appropriate level of accuracy of AI systems and for enabling providers to declare the relevant accuracy metrics and levels."[128] In addition, it requests establishing "appropriate and relevant tools and metrics to measure accuracy against suitably defined levels, that are specific to certain AI systems in consideration of their intended purpose".

Defining Accuracy

In this context, it is notable that the AIA itself does not provide an explicit definition of accuracy. In the context of the trustworthiness of technical systems, accuracy is commonly defined as the "closeness of results of observations, computations, or estimates to the true values or the values accepted as being true."[129] In line with this definition, accuracy is accepted as a central performance metric in the field of machine learning. However, there exist many more criteria to measure the performance of a system, and depending on the technical approach and the purpose of the AI system, accuracy in this narrower sense may represent a more or less suitable performance criterion.[130]

Therefore, it is noteworthy that in its standardisation request to CEN-CENELEC, the Commission declares that accuracy "shall be understood as referring to the capability of the AI system to perform the task for which it has been designed".[131] This definition conceptualises accuracy as a much more general criterion for evaluating the performance of high-risk AI systems. This interpretation differs from the narrower definition of statistical accuracy, as explicitly mentioned by the Commission[132], and goes beyond other definitions provided by standardisation organisations, such as the ISO definition referenced above.

While such a broader definition is more suitable for accommodating the wide range of different AI systems, risk domains, and use cases covered by the AI Act, it introduces additional complexity in achieving consistent and unambiguous operationalisation, as the performance of a system for a certain task can be measured along many different dimensions. For example, accuracy may then not only refer to the closeness of an output to the true value but could also relate to dimensions such as the time taken to

[126] Soler, G., De Nigris, S., Bassani, E., Sanchez, I., Evas, T. André, A. & Boulangé, T. (2024). Harmonised Standards for the European AI Act. JRC 13943 Science for Policy Brief. Available at https://publications.jrc.ec.europa.eu/repository/handle/JRC139430.

[127] European Commission (2023). Commission Implementing Decision on a standardisation request to the European Committee for Standardisation and the European Committee for Electrotechnical Standardisation in support of Union policy on artificial intelligence. Register of Commission - Documents C(2023)3215. Available at https://ec.europa.eu/transparency/documents-register/detail?ref=C(2023)3215&lang=en.

[128] European Commission (2023). Annex I and Annex II to the Commission Implementing Decision on a standardisation request to the European Committee for Standardisation and the European Committee for Electrotechnical Standardisation in support of Union policy on artificial intelligence. Register of Commission - Documents C(2023)3215. Available at https://ec.europa.eu/transparency/documents-register/api/files/C(2023)3215_1/de00000001048943?rendition=false.

[129] ISO/IEC JTC 1 (2022). ISO/IEC TS 5723:2022. Trustworthiness – Vocabulary. 1-9. Available at https://www.iso.org/standard/81608.html.

[130] See, e.g., Hacker, P. (2023). The European AI liability directives–Critique of a half-hearted approach and lessons for the future. *Computer Law & Security Review, 51*, 105871.

[131] European Commission (2023) Annex II, Point 2.6.

[132] European Commission (2023).

perform a task or the completeness of the output. Whether these different performance dimensions matter in a particular context will depend on the use case, the "problem class" addressed by the AI system[133], the technical approach employed, and the associated risks, among other factors.

General Principles and Elements for Standards on Accuracy Specifications for High-Risk AI Systems

Similarly to the transparency requirement, it is not feasible for a uniform standard to adequately define the relevant accuracy criteria, metrics, thresholds, and measurement methodologies for every high-risk AI system covered by the broad scope of the AIA. However, some cross-cutting elements and principles may be included in general standards on accuracy specifications based on existing recommendations, best practices and standardisation efforts. This could be complemented by technology-specific standards on AI accuracy focusing on some types of systems, such as NLP or computer vision, which are currently under development.[134] In addition, as suggested for the transparency requirement, standards targeted to individual risk domains can establish more tangible and specific criteria and thresholds by referencing and evaluating actual use cases and the specific conditions within each domain.

If implementation efforts of the AI Act follow the broad interpretation suggested by the Commission[135], there is an a priori question on the selection of relevant performance dimensions and performance criteria for assessing the accuracy of a high-risk AI system. As a general principle, the selection of performance criteria should focus on those that are most relevant to the potential risks associated with a high-risk AI system.[136] For example, the time it takes for a high-risk AI system to produce an output should be considered a relevant criterion of performance (and thus accuracy under the AIA), if it is relevant for avoiding harm, as is likely in the case with AI systems used in connected cars. In other systems, where slow processing of inputs into outputs does not pose a risk to health, safety, or fundamental rights, the time dimension may not be considered a relevant accuracy criterion. General standards could facilitate effective and efficient implementation, by providing a list of potential performance criteria, making it easier for providers to identify the relevant ones for their specific high-risk AI system.

For any accuracy criterion, numerous metrics can be used to measure it. For example, the accuracy of an AI system (in the narrow sense) that solves a classification problem may be measured by metrics such as accuracy, precision, recall, the F-score, or more complex measures.[137] All of these metrics can provide complementary information and insights into the accuracy of an AI system. Thus, it will often be suitable to report multiple complementary metrics for the measurement of a particular accuracy criterion. As suitable metrics vary across different technical approaches employed in AI systems and the diverse use cases to which they are applied, it is difficult to standardise a uniform set of metrics even when the accuracy criterion is defined. As an intermediate step, standards could provide templates of established

[133] For example, whether the AI system addresses a regression or a classification task; or more generally a supervised learning, unsupervised learning or reinforcement learning problem.

[134] Soler et al. (2024).

[135] European Commission (2023) Annex II.

[136] Thus, the selection of relevant performance metrics can be based on the risk identified as part of the risk management system, Art. 9 (2)(a) AIA.

[137] See, e.g., Hossin, M., & Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process*, *5*(2), 1-11.

and widely recognised metrics (such as those mentioned above for the example of classification problems) to assist providers in selecting appropriate metrics.

In addition to these computational-centric measures, the NIST AI RMF suggests that measures of accuracy should consider human-AI teaming and demonstrate external validity, referring to the ability of AI systems to generalise beyond the training conditions.[138] Hence, if an AI system is supposed to interact with human users, standards on measurement methodologies may define the necessary scope of testing human-AI interactions and their outcomes. To support the consistent evaluation of generalisability, general minimum requirements on the use of separate test or holdout data sets may be specified, such as making these elements mandatory for the testing procedure of high-risk AI systems.

The NIST AI RMF further specifies that measurements of accuracy metrics "should always be paired with clearly defined and realistic test sets – that are representative of conditions of expected use – and details about test methodology".[139] This aligns with the requirements of the AIA regarding training, validation, and testing data.[140] As performance metrics are only informative in the context of the employed testing procedures and test sets, standardisation for testing procedures should build on this general principle. In addition, standards may require that accuracy measurements may include disaggregation of results for different data segments, especially, when this disaggregation can help to assess the performance of the high-risk AI system for relevant subgroups of users and facilitate the identification and mitigation of risks, such as discriminatory outcomes.[141]

Need for Additional Standards on Accuracy Specifications for Individual Risk Domains

As previously discussed in Section 3 regarding standardisation in the context of the transparency requirement under the AIA, general and horizontally applicable standards for implementing the requirements for high-risk AI systems face significant challenges due to the broad scope of the AIA. If standards are too abstract, they can only offer limited guidance and consistency. If standards are overly specific or impose excessively narrow requirements on the selection of accuracy criteria, their measurement, or too strict thresholds, they risk limiting beneficial use cases of high-risk AI systems and stifling innovation. If standards are specific but too lenient, they could undermine the effective protection of health, safety, fundamental rights, and the mitigation of risks. Hence, there is a need for more specific standards, which can also be complemented by guidelines. While more specific standards carry the risk of promoting fragmentation, the AIA already establishes a common framework that could be further supported by general principles and elements outlined in cross-cutting standards (see the discussion above).

This is also emphasised by recent academic studies, making specific reference to the accuracy of AI systems. Based on a systematic review of the literature on evaluation criteria for trustworthy AI,

[138] NIST (2023).

[139] NIST (2023).

[140] Art. 10 (3) and Recital 67 AIA.

[141] As mentioned above, enforcing a specific fairness metric, which are used to determine performance differences between subgroups, carries the risk of inadvertently promoting other biases. This should be carefully considered when establishing guidelines or standards for assessing the performance of AI systems for relevant subgroups.

McCormack and Bendechache highlight that the most suitable metrics for accuracy (as well as for many non-functional properties) will not be the same for different use cases.[142] For example, suitable metrics for accuracy will differ between a medical device and a credit lending algorithm. Therefore, they conclude that "[i]t is not sufficient for AI evaluation criteria and metrics to be developed in a one-size-fits-all model" and that suitable evaluation criteria and acceptable levels for thresholds must be developed for the specific various use cases of a sector.[143]

This is even more relevant when standards are envisioned to specify precise thresholds for accuracy metrics beyond the general principle of appropriateness, which is further discussed in Section 4.2. In the context of related EU legislation, such as the new Product Liability Directive and the proposal for a European AI Liability Directive,[144] standards have been proposed as "safe harbours" offering legal certainty by specifying thresholds and criteria that create a presumption of compliance.[145] Such thresholds for accuracy may be defined under the AIA in line with Art. 40 (1) AIA. However, due to the challenges mentioned above, precise thresholds can rarely be adequately defined on a general horizontal level.

These considerations align with the earlier comments on standards for transparency requirements (see Section 3.2). Thus, also for the implementation of accuracy requirements, standards developed on the level of specific risk domains are viewed as a promising approach to balancing the trade-off between general and consistent but useful and appropriate implementation.

4.2 Towards an Operationalisable Benchmark for Appropriate Accuracy

The AIA requires that high-risk AI systems must achieve an appropriate level of accuracy.[146] Beyond the challenge of defining and measuring accuracy, as discussed in the previous section, this raises the practical question of when a level of accuracy should be deemed appropriate. In this context, the AIA only indicates that an appropriate level of accuracy should be achieved by high-risk AI systems "in light of their intended purpose and in accordance with the generally acknowledged state of the art".[147]

The reference to the generally acknowledged state of the art suggests that appropriate accuracy can be assessed by considering the alternatives commonly available for performing the same task to which a high-risk AI system is intended to be applied. In many contexts where AI systems are expected to operate, the task has previously been performed by either a human or a non-AI technical system. Consequently, a lower bound for the appropriate accuracy of a high-risk AI system (as conceptualised in Section 4.1) is

[142] McCormack, L., & Bendechache, M. (2024). A comprehensive survey and classification of evaluation criteria for trustworthy artificial intelligence. *AI and Ethics*, 1-22.

[143] McCormack & Bendechache (2024), p. 18.

[144] Directive (EU) 2024/2853 of 23 October 2024 on liability for defective products and repealing Council Directive 85/374/EEC, http://data.europa.eu/eli/dir/2024/2853/oj; Proposal for a Directive on adapting non-contractual civil liability rules to artificial intelligence (AI Liability Directive), COM/2022/496 final, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52022PC0496.

[145] Hacker, P. (2023). In addition, or alternatively, quantitative thresholds could also be specified to indicate a presumption of non-compliance.

[146] Art. 15 (1) AIA.

[147] Recital 74 AIA.

determined by the average accuracy achieved by humans or, if a non-AI technical system superior to human accuracy is commonly applied to the task, the accuracy of that system.

Whenever the accuracy for a given task is measurable based on accepted criteria and metrics, this general principle can be readily applied to specific use cases. As high-risk AI systems are presumed to introduce additional risks compared to the alternatives of a human or a non-AI technical system performing the same task (for example, due to their opaqueness),[148] the principle can be justified on the grounds that the AI system should, at minimum, achieve the same accuracy as these existing alternatives. Based on this reasoning, it could be argued that a high-risk AI system might, in fact, need to achieve a sufficient improvement in accuracy over existing alternatives to offset the potential risks it may introduce. However, implementing this principle is challenging, as risks and benefits cannot be simply considered to offset one another due to their multidimensional and often abstract nature. Furthermore, Recital 72 indicates that the state of the art itself should serve as the relevant benchmark for determining appropriate accuracy, rather than requiring an incremental improvement over it.

If other AI systems are already applied to the same task and considered state of the art, the comparison of relevant alternatives should include these AI systems. However, it is then important to not define the state of the art too narrowly. For instance, if a single AI system becomes available that achieves exceptionally higher accuracy compared to all other alternatives, this should not immediately render all other AI systems with lower accuracy inappropriate. Such an approach could undermine previous investments made by deployers in existing AI systems. In addition, fostering competition and innovation requires allowing other providers the opportunity to catch up with technological leaders. Overly narrow requirements regarding the accuracy of AI systems could stifle these incentives and hinder progress.

Therefore, to be recognised as state of the art, the corresponding level of accuracy should have been established over a reasonable period and should not be deemed exceptional. What constitutes a reasonable period may thereby vary across application contexts and depend on the typical lifetime of a product or service within a given application context. Furthermore, following the same general reasoning, the difference in accuracy between a high-risk AI system and an AI system considered state of the art should be sufficiently significant to justify deeming the former’s accuracy inappropriate.

As an exception to this general rule, there could exist high-risk AI systems that provide a significantly lower accuracy than the state of the art, but offer other benefits not available from alternatives considered state of the art. In such cases, it may still be reasonable to allow these high-risk AI systems to operate if the benefit is deemed sufficiently important from a societal perspective. Therefore, the proposed general principle should not be established as an absolute benchmark. However, deviating from it could place the burden of proof on the provider to demonstrate that the benefits outweigh the risks associated with the lower accuracy of the high-risk AI system. To promote innovation, the AI Office or national competent authorities may develop procedures or tools to support these assessments, thereby offering legal certainty to providers. Conversely, compliance with the general principle could offer a presumption of conformity.

[148] Recital 5 AIA.

A further complication to the proposed principle arises when accuracy is difficult to measure. For instance, this may occur when the ground truth for assessing accuracy is unknown (see p. 24). In addition, performance comparisons between humans and AI systems are not always straightforward, as AI systems may perform tasks differently.[149] In some cases, AI systems may be introduced to perform completely new tasks, where no alternative, and thus no state of the art, exists.

In these cases, a possible benchmark for appropriate accuracy could be that the accuracy of the high-risk AI system is sufficient to avoid causing risks to the intended group of users or a specific subgroup of these users. As this benchmark requires significantly more consideration of system-specific and use case-specific contexts, evaluations would need to be conducted on a case-by-case basis. Guidelines or standards tailored to specific risk domains could potentially provide more detailed guidance for such situations.

In general, the accuracy of a high-risk AI system only reflects some form of average performance. Even an AI system with very high accuracy may encounter input instances where it is highly uncertain about the correct output. Therefore, quantifying the uncertainty (or conversely, the confidence) of a high-risk AI system for each individual output during inference, and effectively communicating this information to humans overseeing or using the system, can serve as additional safeguards against risks from inaccurate outputs.[150] Thus, integrating such uncertainty mechanisms and uncertainty-aware explanations into high-risk AI systems should be recognised as effective tools for mitigating the risks associated with their use.

[149] NIST (2023), p. 6.

[150] Hüllermeier & Waegeman (2021); Bobek & Nalepa (2021).

5 Conclusions and Recommendations

This Issue Paper takes a first step toward analysing and deriving recommendations for the efficiency and effective operationalisation of selected requirements under the AIA. The paper identifies open questions about the implementation of provisions for high-risk AI systems, evaluates potential approaches to address these questions in light of the underlying trade-offs, and proposes further steps to establish actionable guidance for providers and deployers of high-risk AI systems. The analysis focuses specifically on the transparency provisions in Art. 13 AIA and the provisions on appropriate accuracy in Art. 15 AIA for high-risk AI systems.

A central challenge to the implementation of the AIA lies in its broad scope. Given the wide range of technical approaches employed in high-risk AI systems and the diverse use cases to which they are applied, deriving general yet useful and appropriate criteria is inherently difficult. To address this challenge, the paper proposes elements and principles that could be established at the general level of the AIA to promote consistency across risk domains and applications. These recommendations include defining main categories of information to be provided under the transparency requirement and establishing a general principle for determining the appropriate level of accuracy of high-risk AI systems. Furthermore, the paper highlights requirements and specifications that demand more granular guidance, such as within specific risk domains, to ensure effectiveness and proportionality. For example, this includes criteria to determine whether the interpretability of a high-risk AI system's outputs requires using intrinsically interpretable models or post-hoc interpretation techniques, as well as guidance on what performance dimensions constitute relevant accuracy criteria for a given high-risk AI system.

Based on an analysis of the AIA provisions, the specific requirements, the technical state of the art in AI systems and methodologies, findings from academic literature, and the involved trade-offs, the following key recommendations are derived:

1. **Implementation of AIA transparency requirements can be operationalised by delineating three main categories of information:**
   a) Information on the characteristics, capabilities and limitations of performance of a high-risk AI system.
   b) Documentation of potential risks, and known or foreseeable circumstances that negatively impact functional or non-functional characteristics of the AI system.
   c) Information and measures to enable interpretability of AI system outputs.

   Transparency about the characteristics, capabilities and limitations of performance of a high-risk AI system can generally also satisfy requirements for the interpretation of AI system outputs. To convey this information, the instructions for use of high-risk AI systems could build on the concept of model cards, which may be augmented to also support the consistent documentation of potential risks and relevant influencing factors. In addition, information about the integration of the AI model into the broader AI system, and how additional data processing at the system level contributes to the AI system's outputs, should be included.

Beyond this general principle, risk-domain-specific analyses may establish the need for intrinsically interpretable models or post-hoc explanation techniques to allow for the interpretability of AI system outputs in specific cases but must account for the limitations of these techniques and the involved trade-offs with other functional and non-functional objectives.

2. **Establishing common practices through guidelines and standardisation for specific risk domains:** Standardisation and the agreed-upon selection of criteria and metrics promote transparency and risk mitigation, as deployers can more easily assess and evaluate provided information. At the same time, this can help providers of AI systems by reducing uncertainty and compliance costs through the adoption of common practices. However, the broad scope of the AIA requirements makes it difficult to set tangible and actionable agreements and standards, while also accounting for the various specificities of the diverse use cases and associated technical approaches underlying AI systems. Hence, guidelines and standards may be developed for specific risk domains of high-risk AI systems to promote common practices for implementation. The classification of high-risk systems specified in Art. 6 (1) and Art. 6 (2) AIA may provide a high-level delineation of these different application domains, although more granular approaches could be necessary in broad domains.

   Guidelines or standards at the level of risk domains can help identify use cases where interpretability of AI system outputs is deemed essential and establish whether intrinsically interpretable models or post-hoc explanation techniques should be considered suitable measures in these specific cases (see Recommendation 1). This requires an analysis of the feasibility and adequacy of these methods in the given context, the trade-offs with other performance and non-performance goals, and their effectiveness in mitigating the specific risks. Additionally, potential challenges to interpretability arising from the interplay between AI models and the system into which they are integrated should be considered.

   Furthermore, establishing common practices for specific risk domains can promote the effective implementation of transparency and accuracy requirements under the AIA by providing guidance on relevant accuracy criteria, metrics, testing procedures, and test sets. Such common practices can complement and support contractual agreements between actors along the AI value chain. This may include specifying commonly accepted templates on these different elements for high-risk providers to choose from for their use cases. In addition, standardised test datasets and testing procedures may be developed with regard to specific use cases and conditions in a particular risk domain. To this end, domain-specific standards should build on established or emerging standards for the transparency and accuracy of AI systems to reduce compliance costs and support global harmonisation.

3. **Timely transparency through post-market monitoring of risks and effective feedback mechanisms across the AI value chain:**
   The analysis of the AIA transparency requirements and their intended purposes highlights the importance of post-market monitoring of risks and effective feedback mechanisms between deployers and providers of high-risk AI systems to ensure timely transparency about risks and to

facilitate cooperative approaches to risk mitigation between providers and deployers of high-risk AI systems.

Such feedback mechanisms should complement mandatory reporting obligations for serious incidents, which must be reported to the market surveillance authorities. In contrast, the envisioned feedback mechanisms for post-market monitoring aim to foster collaboration between providers and deployers to anticipate and mitigate emerging risks, consider limitations of performance and possible remedies that arise with respect to context-specific applications, or identify changes in the inputs and environments that could affect the performance of the AI system. The information gathered through such feedback mechanisms should then be dynamically incorporated into the instructions for use of the concerned high-risk AI system and made accessible to all deployers to promote timely transparency. In turn, effective post-market monitoring of risks can alleviate the burden of ex-ante risk identification and evaluation, which is often constrained by the complex, uncertain, and dynamic contexts of many high-risk AI systems.

4. **Human accuracy and technical state of the art as a general benchmark for the appropriate accuracy of high-risk AI systems:**
   As a general principle across risk domains, appropriate accuracy should be determined based on the state of the art of commonly available alternatives performing the same task. In many contexts where AI systems are expected to operate, the task has previously been performed by either a human or a non-AI technical system. Consequently, a lower bound for the appropriate accuracy of a high-risk AI system is determined by the average accuracy achieved by humans or, if a non-AI technical system that is superior to human accuracy is commonly applied to the task, by the accuracy of that system.

   If other AI systems are already applied to the same task and considered state of the art, these systems should be included in the comparison of relevant alternatives. However, it is then important not to define the state of the art too narrowly, as this could undermine predictability, competition, and innovation. Therefore, to qualify as state of the art, the corresponding level of accuracy should have been established over a reasonable period of time and should not be considered exceptional.

   Compliance with this proposed principle could establish a presumption of conformity. As an exception, high-risk AI systems may be permitted to operate even if they fail to meet the general principle, but only if the benefits provided by such systems are deemed sufficiently important from a societal perspective. In such cases, the burden of proof should be placed on the provider to demonstrate that the benefits of the high-risk AI system outweigh the risks associated with its lower accuracy. To promote innovation, the AI Office or national competent authorities may develop procedures or tools to support these assessments, thereby offering legal certainty to providers.

   Whenever the accuracy for a given task is not measurable using accepted criteria and metrics, a possible benchmark for appropriate accuracy could be that the accuracy of the high-risk AI system is sufficient to avoid causing risks to the intended group of users or a specific subgroup of these users. As this benchmark requires significantly more consideration of system-specific and use case-specific contexts, evaluations would need to be conducted on a case-by-case basis.

cerre Centre on Regulation in Europe

Avenue Louise 475 (box 10)
1050 Brussels, Belgium
+32 2 230 83 60
info@cerre.eu
www.cerre.eu

Centre on Regulation in Europe (CERRE)
CERRE Think Tank