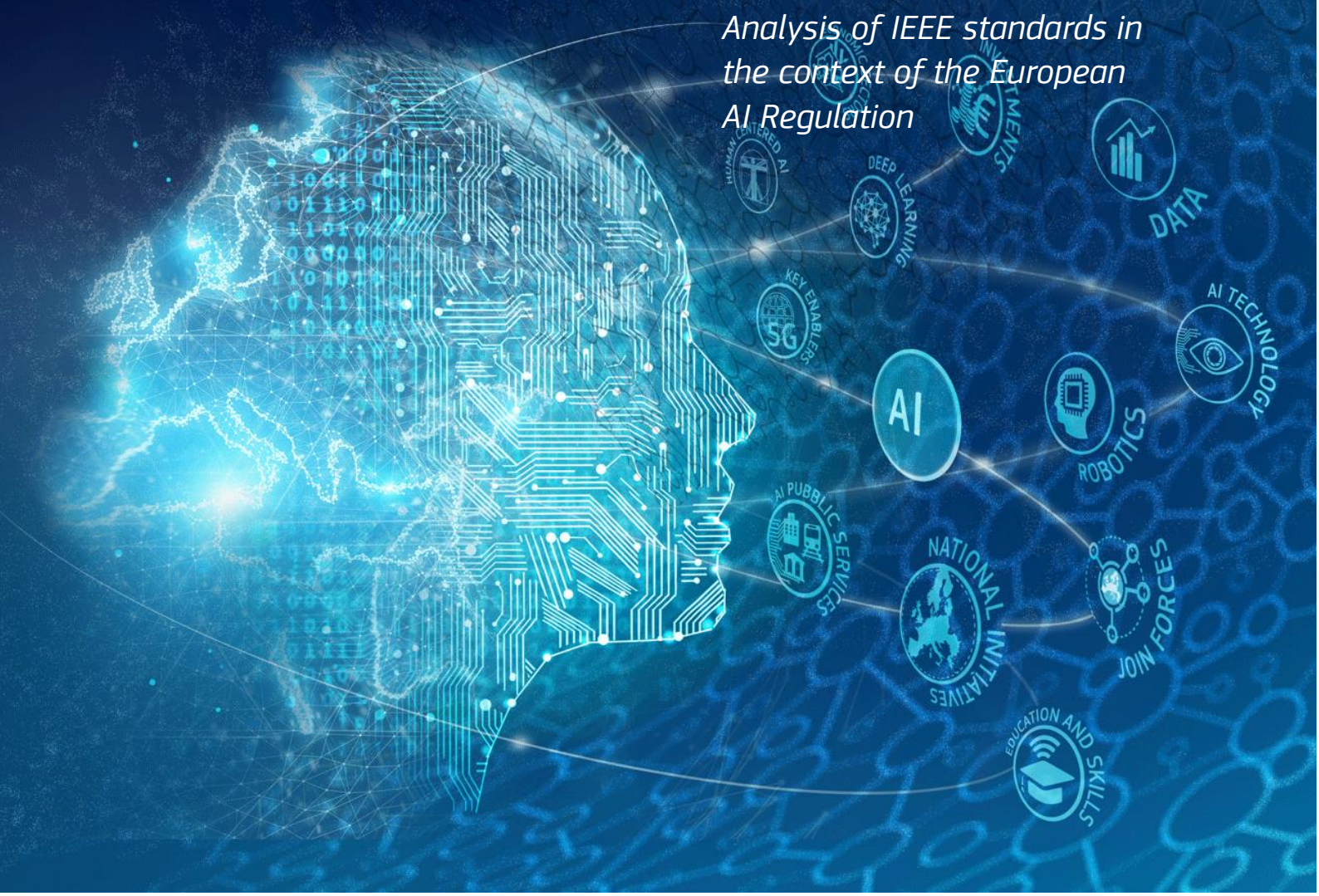
European Commission

JRC TECHNICAL REPORT

AI Watch: Artificial Intelligence Standardisation Landscape Update

Analysis of IEEE standards in the context of the European AI Regulation



Joint Research Centre

EUR 31343 EN

Contact information
Name: Josep Soler Garrido
Address: Edificio EXPO, Avda Inca Garcilaso sn, 41092 Seville, Spain
Email: josep.soler-garrido@ec.europa.eu

EU Science Hub
https://joint-research-centre.ec.europa.eu

JRC131155

EUR 31343 EN

PDF ISBN 978-92-76-60450-1 ISSN 1831-9424 doi:10.2760/131984 KJ-NA-31-343-EN-N

Luxembourg: Publications Office of the European Union, 2023

How to cite this report: Soler Garrido, J., Tolan, S., Hupont Torres, I., Fernandez Llorca, D., Charisi, V., Gomez Gutierrez, E., Junklewitz, H., Hamon, R., Fano Yela, D. and Panigutti, C., *AI Watch: Artificial Intelligence Standardisation Landscape Update*, Publications Office of the European Union, Luxembourg, 2023, doi:10.2760/131984, JRC131155.

Contents

Abstract

The European Commission presented in April 2021 the AI Act, its proposed legislative framework for Artificial Intelligence, which sets the necessary regulatory conditions for the adoption of trustworthy AI practices in the European Union. The AI Act adopts a risk-based approach, laying down a set of legal requirements for certain AI systems, primarily those that are classified as high-risk. At the time of writing this report, the AI Act is under negotiation between the co-legislators, the European Parliament and the Council of the European Union. Once an agreement is found and the final legal text comes into force, standards will play a fundamental role in supporting providers of concerned AI systems. Standards are set to bring the necessary level of technical detail into the essential requirements prescribed in the legal text, defining concrete processes, methods and techniques that AI providers can implement in order to comply with their legal obligations. Indeed, harmonised standards —produced by European standardisation organisations based on a formal standardisation request issued by the European Commission— provide operators with a presumption of conformity with the legal requirements of the EU harmonisation legislation in question once they are accepted and the reference of those standards is published in the Official Journal. However, the drafting of standards is an elaborate process, requiring extensive technical expertise and the coordination of multiple stakeholders. Fortunately, AI has been an active area of work by many standards development organizations in recent years, resulting in a wealth of specifications with the potential to support the future AI Act. In this report, we analyse a set of such specifications, selected from the broad range of standards and certification criteria produced by the IEEE Standards Association covering aspects of trustworthy AI. Several of the documents analysed have been found to provide highly relevant technical content from the point of view of the AI Act. Furthermore, some of them cover important standardization gaps identified in previous analyses. This work is intended to provide independent input to European and international standardisers currently planning AI standardisation activities in support of the regulatory needs. This report identifies concrete elements in IEEE standards and certification criteria that could fulfil standardisation needs emerging from the European AI Regulation proposal, and provides recommendations for their potential adoption and development in this direction.

Foreword

This report is published in the context of AI Watch, the European Commission knowledge service to monitor the development, uptake and impact of Artificial Intelligence (AI) for Europe, launched in December 2018.

AI has become an area of strategic importance with potential to be a key driver of economic development. AI also has a wide range of potential social implications. As part of its Digital Single Market Strategy, the European Commission put forward in April 2018 a European strategy on AI in its Communication "Artificial Intelligence for Europe". The aims of the European AI strategy announced in the communication are:

- to boost the European Union's technological and industrial capacity and AI uptake across the economy, both by the private and public sectors;
- to prepare for socio-economic changes brought about by AI; and
- to ensure an appropriate ethical and legal framework.

In December 2018, the European Commission (EC) and the Member States published a "Coordinated Plan on Artificial Intelligence", on the development of AI in the EU. The Coordinated Plan already mentioned the role of AI Watch to monitor its implementation.

Subsequently, in February 2020, the Commission unveiled its vision for a digital transformation that works for everyone. The Commission presented a White Paper proposing a framework for trustworthy AI based on excellence and trust.

These efforts were substantiated in April 2021, when the EC proposed a set of actions to boost excellence in AI, and rules to ensure that the technology is trustworthy. The proposed Regulation on a European Approach for Artificial Intelligence (short the "AI Act") and the update of the Coordinated Plan on AI aim to guarantee the safety and fundamental rights of people and businesses, while strengthening investment and innovation across EU countries. The 2021 review of the Coordinated Plan on AI refers to AI Watch reports and confirms the role of AI Watch to support implementation and monitoring of the Coordinated Plan.

AI Watch monitors the European Union's industrial, technological and research capacity in AI; AI-related policy initiatives in the Member States; uptake and technical developments of AI; and AI impact. AI Watch has a European focus within the global landscape. In the context of AI Watch, the Commission works in coordination with Member States. AI Watch results and analyses are published on the AI Watch Portal.

From AI Watch's in-depth analyses, we will be able to better understand the European Union's areas of strength and the areas where investment is needed. AI Watch will provide an independent assessment of the impacts and benefits of AI on growth, jobs, education, and society.

AI Watch is developed by the Joint Research Centre (JRC) of the European Commission in collaboration with the Directorate-General for Communications Networks, Content and Technology (DG CNECT).

This report addresses the following objective of AI Watch: to monitor and gather information on the work on Artificial Intelligence of European and international Standards Development Organizations (SDOs), and assess their relevance in the context of the European regulatory framework proposal on Artificial Intelligence.

Acknowledgements

The authors would like to thank the IEEE Standards Association for their engagement in many useful and productive discussions on AI standardization with the European Commission, and especially for making this JRC report possible by facilitating the necessary documents and standards, many of which were still in draft form when reviewed. Similarly, they would like to thank their colleagues from DG CNECT Salvatore Scalzo, Tatjana Evas, Filipe Jones Mourao, Thierry Boulange and Antoine-Alexandre Andre for their continued support and especially for the excellent teamwork. In addition, they would like to extend their gratitude to the many colleagues from various European Commission services involved on AI standardisation activities and supporting our work, especially Antonio Conte (DG GROW), Thomas Reibe (DG CNECT) and Emilio Dávila (DG CNECT). In a similar way, they would like to thank the European Standardisation Organizations for the many valuable discussions, and especially to the participants in AI standardisation roadmap definition activities within the CEN/CENELEC JTC21 Strategic Advisory Group. Last but not least, they would like to thank the many individuals that have been involved in AI Watch, and in particular Stefano Nativi and Sarah de Nigris for their valuable contributions on AI standardisation, as well as Eva Martínez and Paul Desruelle for their leadership and support.

Authors

Soler Garrido, Josep

Tolan, Songül

Hupont Torres, Isabelle

Fernández Llorca, David

Charisi, Vicky

Gómez Gutiérrez, Emilia

Junklewitz, Henrik

Hamon, Ronan

Fano Yela, Delia

Panigutti, Cecilia

Executive Summary

The European Commission's proposal for the regulation of Artificial Intelligence in the European Union –the "AI Act"– is, at the time of publishing this report, under negotiation by the European Parliament and Council. Recent progress made by the two co-legislators, notably with the adoption by the Council of its general approach in December 2022, paves to the way to the upcoming start of inter-institutional negotiations and the final stages of the legislative process. Once the regulation enters into force, and after a transitional period, AI systems classified as high-risk will be required to comply with a comprehensive set of AI trustworthiness requirements prior to their placing on the market or putting into service. In line with the EU legislative framework on products, harmonised standards will play a key role in facilitating the placement of AI products on the market by defining technical solutions to fulfil those requirements, and providing relevant AI system providers with a legal presumption of conformity. Given the importance of standards to ensure the effectiveness of the AI regulation, the European Commission has already started the process to adopt a standardisation request providing a formal mandate to European Standardisation Organisations to develop the necessary standards. The Commission's standardisation request is expected to be formally adopted in early 2023, marking the start of a period of four months during which the standardisation bodies addressed by the request should prepare and submit a work programme for the provision of the standardisation deliverables requested. This report is mainly intended as an input to the European Standardisation Organizations to support the development of this work programme and the planning of their activities.

As highlighted in the standardisation request itself, the set of technical specifications to be adopted in support of the AI act is expected to rely significantly on international work. This is mostly expected to come from ISO and IEC, given the relevance of that work and existing agreements in place that facilitate its adoption in the European context. However, relevant AI standardisation work has also been carried out by other international standardisation bodies, such as the IEEE Standards Association. In this report, we present a systematic analysis of a set of 8 IEEE documents. These include standards from the IEEE 7000 series for ethically aligned autonomous and intelligent systems as well as selected suites of certification criteria from the IEEE ethics certification program for autonomous and intelligent systems. The analysis of these standards has resulted in the identification of valuable content towards operationalizing requirements related to AI bias, human oversight, record keeping and risk management. Similarly, the analysis of the certification criteria indicates that they could in the future provide a basis to fulfil the need for implementable methods for verifying compliance with the AI regulation. In light of this, we present an in-depth examination of the content of IEEE standards, providing a comparison with existing ISO/IEC work and identifying areas that may require adaptation to European needs, in order to facilitate their potential integration within European standardisation work for the AI Act.

Important observations and recommendations deriving from our analysis include the following:

- The IEEE P7003 Draft Standard for Algorithmic Bias Considerations should be considered a relevant source of technical specification for the operationalization of AI Act requirements in relation to bias. Unwanted bias is a main cause of many of the potential risks of AI systems, and the IEEE document provides a comprehensive coverage of this key trustworthiness aspect. In addition, at the moment, a technical specification from ISO/IEC appears to be at a considerably less advanced stage of development. Considering this, we recommend to explore formulas for leveraging IEEE work in the European context. AI bias is strongly linked to all of the trustworthiness requirements for high-risk AI systems in the legal text. Therefore, a work item currently planned by CEN-CENELEC JTC21 on an overarching unified approach on trustworthiness characteristics could represent a suitable integration instrument for this work.
- The IEEE P7001 Draft Standard for Transparency of Autonomous Systems provides relevant coverage of human oversight aspects. Measures described in this document for enabling understanding of the AI system's function by its users and operators are highly relevant. The document covers various levels of transparency, with increasing degrees of sophistication. In the context of the AI Act, basic approaches that are mature, robust and effective are particularly suitable. On the other hand, some of the highest transparency levels defined in the document demand explanations of system decisions that may be difficult to achieve in practice, or require the use of experimental research methods potentially not ready for harmonised standardisation. These advanced transparency levels are not strictly required for compliance with human oversight requirements in the AI Act, and the standardisation request explicitly asks for methods that are consolidated and robust. Our analysis looks at this IEEE document as well as ongoing ISO/IEC work related to human oversight, and highlight the need for European standards to consider technological

maturity and known limitations of the different methodologies available, clearly outlining the skills required by stakeholders making use of them. Some of the content in IEEE P7001 can contribute to this effort, as detailed in this report.

- The IEEE P7001 draft standard also provides valuable coverage of record-keeping requirements in the European AI regulation proposal. Most relevant is the content related to transparency towards expert stakeholders facilitating the inspection and investigation of incidents involving AI systems. This includes elements that would assist AI providers in implementing effective logging and record-keeping measures, including specific guidance on recording AI system decisions, intermediate states and internal events. This may complement existing standards covering concerns more generally applicable to all software systems, such as logging formats. Considering this, we recommend to explore ways to integrate some of this IEEE work in the work programme for European AI standardisation, possibly also within the planned CEN-CENELEC JTC21 standard on an overarching unified approach on trustworthiness characteristics.
- The IEEE 7000 Standard Model Process for Addressing Ethical Concerns during System Design is a useful reference towards operationalising risk management requirements in the AI Act. In contrast to other international standards reviewed, including those from ISO/IEC, we highlight its level of prescriptiveness and product orientation, as it provides a process to systematically consider and address ethical values and risks in the design of an AI system, translating them into traceable product requirements. Another positive aspect of this standard is its ongoing adoption at ISO/IEC level, which would facilitate subsequent adoption at European level. However, this standard cannot by itself cover all the elements contained in the risk management article of the AI act. For example, it would have to be expanded to provide more thorough coverage of the entire AI lifecycle beyond the initial design stages, including development, deployment, operation and post-market monitoring. Furthermore, tailoring to the specific risks and European values at the core of the EU regulation proposal would have to be built into the standard and prioritised in conformity checks. In addition, being a process-oriented standard, it should be complemented with concrete specification and guidance at the technical level, for example as currently being planned within CEN-CENELEC JTC21 work on AI systems risk catalogue and risk management. Indeed, we see standardisation work at the European level as an instrument for integration and supplementation of valuable content from ISO/IEC and IEEE, preventing fragmentation when it comes to standards on AI risk management, hence facilitating the work of AI providers to achieve compliance.

In conclusion, our analysis has identified concrete elements in IEEE standards that deserve consideration by European standardisers preparing technical specifications in support of the European AI regulation. In many cases, the specifications analysed are either at a more advanced stage compared to equivalent ISO/IEC standards, or highly complementary. Therefore, we encourage European standardisers to explore formulas to fully leverage existing international work, not only from ISO/IEC but also from IEEE, in the work plan to be prepared in response to the upcoming standardisation request from the European Commission, complementing it where necessary according to European specificities.

1 Introduction

In April 2021, the European Commission presented its proposal for the regulation of Artificial Intelligence (AI), the "AI Act" [1], with the objective to set the conditions for the development and use of trustworthy AI practices in the European Union. The AI Act follows a risk-based approach, defining a set of obligations for providers of AI systems depending on their risk profile. Providers of high-risk AI systems (including those systems which pose risks to fundamental rights, health and/or safety of humans) will be required to address concrete requirements defined in Title III, Chapter 2 of the legal text (Articles 8 to 15); these requirements relate to risk management, data and data governance, technical documentation, record-keeping, transparency and provision of information to users, human oversight, accuracy, robustness and cybersecurity.

The legal text does not, however, specify how to fulfil these requirements at the technical level. Instead, as is the case with European regulations following the New Legislative Framework [2], it defines the essential, high-level requirements to protect public interests, and foresees the creation of European harmonized standards needed for products to conform to these requirements. In this context, the AI Act will be supported by a series of technical specifications produced by European Standardisation Organisations[1] (ESOs) [3] by the time it enters into force. These specifications, while being of voluntary nature, provide presumption of conformity with the legal requirements, and hence play a fundamental role in ensuring a level playing field for all AI providers, regardless of their size and resources, as well as in simplifying the conformity assessment procedure. The process of defining AI standards and technical specifications in support of the AI Act does not start from scratch. The ESOs are able to leverage existing standards and technical specifications, notably through cooperation agreements with international standardization organisations, such as the Vienna agreement [4] between CEN and ISO or the Frankfurt agreement [5] between CENELEC with IEC. The adoption of existing international work is the most efficient way to avoid duplication of work and to greatly reduce the time needed to develop the wide range of standards needed for the upcoming AI regulation.

In this context, a crucial first step in the process of defining a European AI standardisation roadmap is to take stock of the existing landscape at the international level. To support this effort, the JRC presented a first AI standardisation landscape analysis in 2021 [6]. This analysis, building on relevant work such as the AI standardisation survey produced by the StandICT.eu project [7], covered approximately 140 standards and standardisation deliverables (such as technical reports, technical specifications and certification criteria) from major international and European Standards Development Organizations (SDOs): ISO/IEC, ETSI, IEEE and ITU-T. A subset of these standards, mainly those relevant in the context of the AI Act that were available in final or draft form at the time, was reviewed in more detail, resulting in a short list of promising standards with the potential to operationalize certain AI Act requirements, as well as a preliminary list of potential standardization gaps. Since the publication of that report, the state of play in the European AI regulatory process has advanced substantially. Significant progress in the negotiation of the legal text has been made by the European Parliament as well as the Council. In parallel, the European Commission is preparing a first standardisation mandate to be sent to the ESOs [8]. This initial mandate, requesting European standards covering the main technical areas underlying the requirements of the AI Act proposal and serving to prepare the technical ground for future harmonised standards, is expected to be adopted in early 2023. After that, it is likely to be subject to updates in order to reflect the outcome of the negotiations between the co-legislators. Despite this, the initial mandate will, once accepted by the ESOs, formalize the considerable task ahead to produce a comprehensive set of specifications supporting the implementation of the AI Act.

In anticipation of the standardisation request, the ESOs have engaged in preliminary roadmap definition activities, and started the process to set up new ad-hoc groups to explore the development of new specifications addressing some of the identified standardisation needs, e.g. on technical aspects of AI trustworthiness or AI system risk management. The European Commission is supporting this process, including through the technical analysis of standards from the point of view of the requirements in the legal text. This report is primarily meant to serve as an input to these planning activities. Despite the considerable amount of ISO/IEC standardisation activities currently underway in JTC1 SC42 on Artificial Intelligence [9], the gaps that need to be addressed in order to fully cover the requirements in the AI Act are considerable. We anticipate that a detailed analysis of existing specifications by other prominent international SDOs could result in the identification of suitable standards that address some of these gaps. In this study, we complement our standardisation landscape analysis by considering standards from the IEEE Standards Association [10].

[1] ESOs are the European Committee for Standardisation (CEN), the European Committee for Electrotechnical Standardisation (CENELEC) and the European Telecommunications Standards Institute (ETSI).

2 Documents in scope

As discussed in the previous section, IEEE standards and related standardisation-related deliverables have already been partially covered in the first AI Watch standardisation landscape analysis presented in 2021. Indeed, a broad range of IEEE standards were included in the collection of approximately 140 documents from various SDOs. This first-level analysis was based on publicly available information (e.g., abstracts and metadata), resulting in an initial mapping to relevant articles in the AI Act proposal. However, given that complete IEEE documents were not available at the time, they were not subject to a more detailed content review, as was the case for standards from other SDOs such as ISO. With this report, we start a more detailed discussion of IEEE standards on Artificial Intelligence and their suitability to operationalize at the technical level the obligations of providers of high-risk AI under the future European AI regulation.

The broad set of documents of interest remains unchanged compared to the previous study, including those coming primarily from two concrete collections: the 7000 series of standards and the Ethics Certification Program for Autonomous and Intelligent Systems (ECPAIS). The documents in the 7000 series address specific concerns at the intersection of technology and ethics with a strong focus on Autonomous and Intelligent Systems (A/IS), and are therefore very relevant references in the context of the human-centred view of the European AI regulation proposal and its focus on risks to fundamental rights. The ECPAIS certification suites are a completely different type of deliverables, complementing process-oriented standards by providing outcome-based criteria to measure key aspects of trustworthy AI such as accountability, transparency and reduction in algorithmic bias, enabling certification of A/IS products, systems, and services against these criteria.

This report presents a detailed analysis of a selected subset of documents from these families, listed in Table 1. It should be noted that, at the time of writing, most of these documents were still work in progress items or not publicly available but were shared by IEEE for this work. This analysis may be extended in future reports in order to cover additional documents, either from the same series or from other relevant families of standards potentially in scope. In the case of IEEE, this may include selected documents from the 2800 series focusing on AI governance and licensing topics, on specific technologies such as deep learning or federated learning, or even covering concrete application sectors of interest in the context of the AI Act, such as healthcare or robotics.

Document	Type	Date
IEEE 7000 Standard Model Process for Addressing Ethical Concerns during System Design	Standard	June 2021
IEEE P7001/D4 Draft Standard for Transparency of Autonomous Systems	Standard	October 2021
IEEE P7003/D1 Draft Standard for Algorithmic Bias Considerations	Standard	January 2022
IEEE 7010 Recommended Practice for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-Being	Standard	March 2020
IEEE P2841 - Framework and Process for Deep Learning Evaluation	Standard	June 2021
IEEE ECPAIS: Accountability Certification Requirements	Certification Criteria	December 2019
IEEE ECPAIS: Transparency Certification Requirements	Certification Criteria	December 2019
IEEE ECPAIS: Bias Certification Requirements	Certification Criteria	December 2019

Table 1. List of analysed IEEE standards and standardisation deliverables.

3 Methodology

A methodology based on expert analysis was implemented to review the standards and certification criteria in scope. This analysis was carried out in two steps:

1) a stand-alone review of individual IEEE standards and certification criteria from the point of view of the requirements of the European AI regulation, and
2) the analysis of complementarities between the ISO/IEC AI standardisation landscape and the IEEE standards perceived as most relevant in the previous step.

For the first step, a pool of JRC experts, all of them authoring this report, was assigned standards from Table 1 for detailed review following a common assessment methodology described in this section. All participants have an in-depth knowledge of the AI regulation and are technical experts on different aspects of trustworthiness of Artificial Intelligence. Each document was assigned to two experts for an independent assessment. Individual reviews were based on a custom-made questionnaire designed to prompt the expert to reflect on how the document fulfils a series of relevant criteria, guiding the process and establishing a common protocol for all the participants in order to promote objectivity and reproducibility in the results.

The considerations contained in the questionnaire are described in Table 2 in detail for the case of standard reviews, with analogous considerations being made in the case of certification criteria.

	Criteria	**Possible considerations**
General Criteria	Artificial Intelligence coverage	Is the content of the standard specific to Artificial Intelligence? Does it at least cover AI broadly, considering the definition of AI in the legal text? Does it focus on specific types of AI methods?
	Domain generality	Is the certification criteria applicable to all high-risk AI systems defined in the AI Act? Are parts of the standard only applicable to specific domains or use cases?
	Maturity and level of detail	How mature and complete is the specification, and considering the technology, process or methodology described, is the level of detail sufficient, clear and of practical use from the point of view of a provider of a high-risk AI system considered in the AI Act. Where would additional detail or guidance be beneficial?
	Compliance management	How accessible is the standard in terms of difficulty for a high-risk AI provider to achieve and demonstrate compliance with it, or for a conformity assessment body to test compliance with it.
	Gaps and complementarities	Have any gaps been identified? For example, regarding coverage of the main legal requirements addressed by the standard, the types of risks addressed, or the AI lifecycle stages?
	Fit within standardisation landscape	How does this standard fit with known standards on artificial intelligence, especially from ISO/IEC, and what is its potential to fill known standardisation gaps from the point of view of the AI Act.
Coverage Analysis	Detailed Requirement Coverage	For the various paragraphs of standardisation-relevant articles in the AI Act describing requirements for high-risk AI systems and obligations of their providers, which clauses and sections of the standard appear to address them? Include an approximate quantitative analysis and any relevant comments in terms of the depth with which they are covered at the technical level.

Table 2. Summary of criteria and considerations defined to guide expert review of standards.

The form initially considers general criteria, including desirable attributes and qualities (AI coverage, domain generality, maturity, technical detail, compliance management), as well as any potential gaps and complementarities needed from the angle of the requirements in the AI Act.

General criteria also include an assessment of the relevance of the respective standard considering the overall AI standardisation landscape. In general, documents that cover standardisation needs not currently addressed by other SDOs are favoured. A preliminary analysis of potential standardisation gaps is available in [6]. However, in this report, we consider a more up-to-date assessment of standardisation gaps guided by ongoing discussions in the context of roadmap definition activities by the ESOs, notably within CEN-CENELEC JTC 21 Strategic Advisory Group. In the context of this group, somewhat stricter criteria are used to consider standards as suitable to operationalize the AI Act requirements, resulting in additional gaps compared to the analysis in [6]. For instance, documents without normative content such as technical reports are not considered, as are documents in early stages of development that may not reach maturity before the regulation comes into force. Considering this, mature and technically sound standards on trustworthy AI aspects covering relevant product requirements, e.g. transparency, addressing bias, human oversight (including explainability and controllability aspects) and robustness are of special interest in order to address existing gaps.

Besides the general criteria described, the form also prompts the reviewer to perform a detailed analysis of the coverage of the individual requirements in the AI Act that the standard provides. In scope are those requirements for high-risk AI systems in Articles 8-15 in Chapter 2 of the legal text which are expected to be included in the initial standardisation mandate from the European Commission. Additionally, some obligations of the providers of high-risk AI systems, namely putting in place a quality management system (Chapter 3, Article 17), are also considered. In line with previous standard analysis activities, other considerations, notably the need for specifications related to conformity assessment, are not part of our analysis.

Once guided analysis of a given standard or certification suite is independently completed, a consolidation meeting takes place where the two assigned experts share their opinions and reach a consensus. After this, a review form is drafted containing both a quantitative as well as a qualitative analysis of the different criteria. Forms for all the reviewed documents are presented in section 4.

When all the reviews of individual standards are finalised and consolidated, those deemed as most relevant –especially regarding their fit within the standardisation landscape– are selected for a second round of analysis by a different set of experts. This second round no longer analyses the standards in isolation, but in the context of relevant AI standardisation work from ISO/IEC. A qualitative analysis is carried out looking specifically for complementarities between individual IEEE standards and the most closely related ISO/IEC specifications, with a focus on those that the European Standardisation Organizations are currently considering as candidates for adoption in the context of the European AI Act. This analysis is presented in section 5.

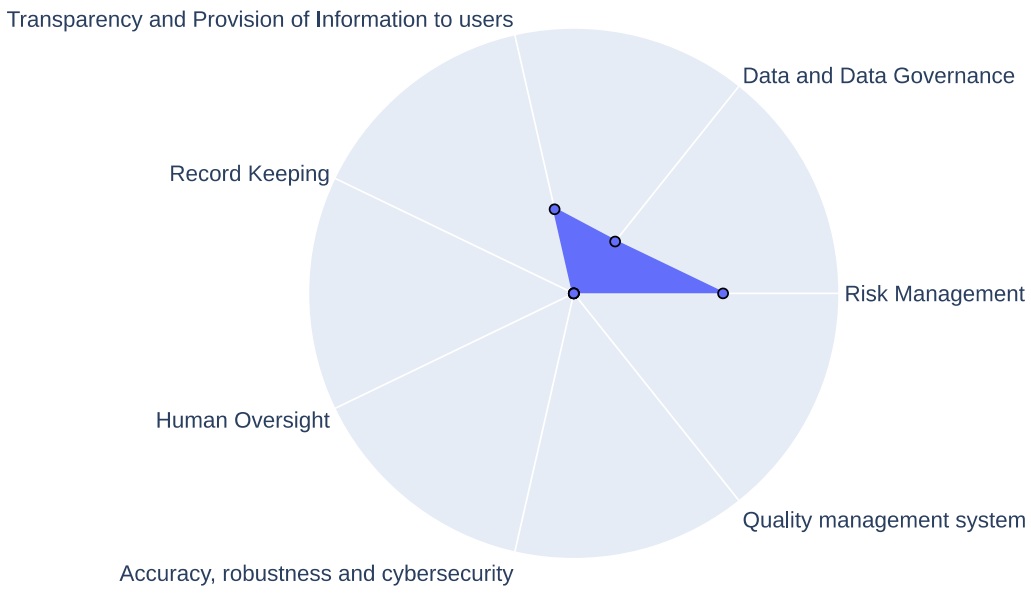
4 Detailed Document Analysis

This section presents the review forms for the analysed documents, covering a qualitative and quantitative analysis based on the criteria described in Table 2. A colour code is used to rate the items under general criteria, with a green circle (●) indicating a strong match with the needs of the AI Act, an orange square (■) indicating a partial match, and a red triangle (▲) being used in cases where potential misalignments are identified which would need to be addressed if the respective specification were to be adopted in the European context to support the implementation of the AI Act. Regarding the detailed coverage analysis, a numeric score up to 1 is derived qualitatively through discussion of the expert reviewers for each requirement in scope. The score can be interpreted as an approximation of the product of breadth and depth of coverage of the various paragraphs in the respective articles of the AI Act. Any score greater than 0 indicates that certain clauses or sections in the standard would contribute to the operationalisation of concrete requirements in the AI Act, with a score of 1 indicating that the standard fully covers the respective article with an optimal level of detail. It should be noted that the ratings and scores presented are only intended to provide an indicative measure of their alignment with the specific standardisation needs of the European AI regulation. Considering that many of the documents reviewed pre-date the presentation of the AI Act proposal, any cases where substantial alignment is found should be considered as very positive.

IEEE 7000 - Standard Model Process for Addressing Ethical Concerns during System Design		
Summary	Relevant standard in the context of Article 9 of the AI Act covering risk management. This document specifies processes that could effectively support providers of high-risk AI systems in designing systems with explicit consideration of individual and societal ethical values, such as fairness, transparency, accountability, sustainability and privacy, addressing relevant risks in the context of the legal text and avoiding potential harms to individuals. The standard covers the concept exploration and development stages, allowing AI providers to systematically explore how the product is expected to perform and operate, consulting with stakeholders in order to elicit their needs and ensure relevant ethical values are reflected and prioritised. These values are taken on board in the design through the definition of ethical requirements and their translation into technical system controls and risk mitigations. Therefore, this standard has been evaluated from the point of view of risk management requirements. In addition, the requirements of transparency and data quality and management are also prominently featured in the standard, and are included in the analysis presented. It should be noted, however, that even if not directly covered in the standard, the controls and mitigations resulting from its application have the potential to address other legal requirements, including, for example, robustness or human oversight measures.	
Artificial Intelligence coverage	●	This standard describes generic processes and is therefore applicable to all kinds of products and services, including AI systems, which are explicitly covered in some sections of the standard. The processes described are best applied to systems where the deployment context is well defined, as opposed to generic products (e.g. general purpose AI systems). This is in alignment with the needs of the AI Act, which demands consideration of the intended purpose when addressing AI risks.
Domain generality	●	This standard can be applied to AI products and services in a horizontal manner, independent of the sector or domain where they operate.
Maturity and level of detail	■	It is a mature specification which has already seen adoption in practice. The processes to be implemented by system designers are described in detail. The scope of this standard is broad, defining a generic approach that is applicable to a wide range of systems. From the point of view of the AI Act, while considering that certain margin of manoeuvre is needed by AI system providers when implementing a risk-management process, more details on how to tailor this standard to AI systems may be beneficial. For example, risks to AI related to the use of data to train algorithms or the opacity of certain machine learning systems are not discussed in detail. Adopters of this standard may welcome guidance on using the processes described to address these or other AI-specific risks, define suitable requirements, translate them into specific technical system controls and assess their effectiveness and trade-offs involved.

Compliance management		Compliance with the provisions of this standard can be demonstrated and verified in an unproblematic manner, as it specifies the necessary evidence in detail. It demands traceability of the ethical requirements embedded in the design, supporting analysis and scrutiny. Despite this, successful application of this standard does not guarantee that the ethical criteria implemented match the specific risks that the AI regulation addresses. The processes defined in this standard are generic and can accommodate different sets of ethical values, leaving the assessment of the concrete ethical criteria employed by the adopters out of scope. The standard does explicitly require that legal requirements should be prioritised. However, in the context of ensuring the risks to fundamental rights considered in the AI Act are properly addressed, detailed assessment of the criteria used by adopters of this standard would be required.
Gaps and complementarities		Certain gaps would need to be addressed (by this or other standards) in order to fully cover risk management requirements in the legal text. The AI Act requires an iterative risk management process spanning the whole lifecycle of AI. Currently, this standard covers primarily the concept exploration and development stages. The processes described are very valuable for AI providers in that they ensure that ethical requirement are embedded in their designs from the early stages. However, the scope of this standard may need to be complemented to cover later stages more prominently (e.g. implementation, deployment and monitoring) for it to effectively define an iterative risk management process able to identify and address new risks throughout the product lifecycle. Finally, as aforementioned, another useful complementarity would be in the form of specifications describing in a more granular and technical level concrete risks, controls and mitigations for AI systems.
Fit within standardisation landscape		Despite the availability of other process-oriented AI standards addressing risk management, the process defined by this document to derive ethical requirements make it a valuable complement. This specification can be integrated with existing processes and practices, e.g. system or software engineering methods and lifecycle models, and therefore it is compatible with other complementary standards.

Requirement Coverage



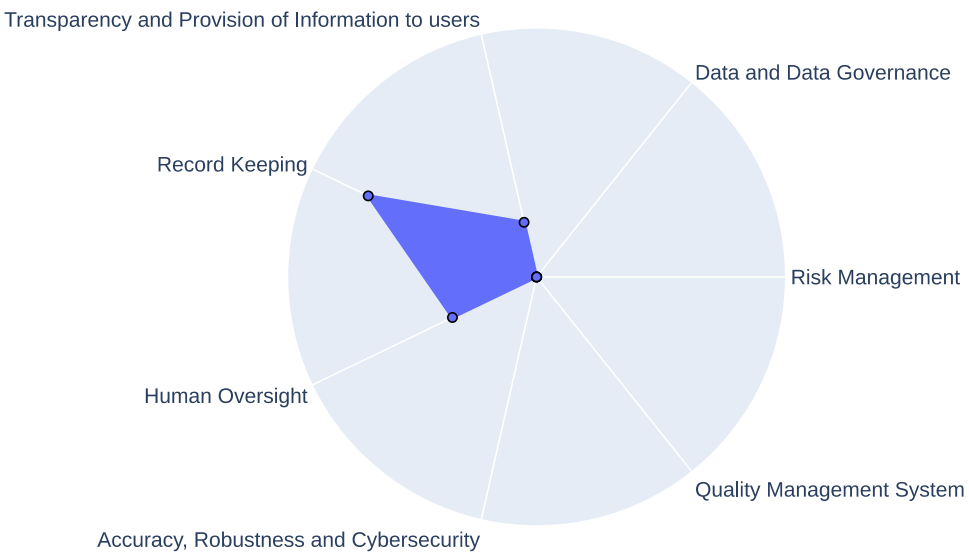
Risk Management	0.56	Many of the requirements related to risk management are meaningfully covered by this standard. Relevant clauses can be found to support the setup of an effective risk management process for AI systems, including: consultation with stakeholders to elicit ethical issues, potential harms and benefits; taking into account the intended purpose of the high-risk AI system; prioritising values and requirements; addressing identified risks systematically; and provision of appropriate information to the users. A notable missing consideration is the definition of an iterative process covering the entire lifecycle and considering post-market monitoring.
Data and Data Governance	0.25	Informative sections of the standard consider data quality as a "control" for AI systems, listing useful considerations regarding accuracy, timeliness, consistency and completeness of data, and highlighting the need to monitor and, to the extent possible, avoid biases. In general, relevant data governance processes may be derived from this standard when applied to AI systems, even if more technical detail and further guidance may be beneficial for operators adopting this standard.
Transparency and Provision of Information to Users	0.33	The standard defines a transparency management process that supports communication to users (as well as other stakeholders) about potential risks of AI systems as well as the concrete measures that the provider has taken to identify and address ethical concerns during system design. Some considerations in the informative sections of the document explicitly cover relevant AI transparency concerns, e.g. explaining algorithm's logic for lay users. Adoption of this standard should help organizations in effectively communicating limitations of AI systems.

Table 3. Analysis of IEEE 7000 - Standard model process for addressing ethical concerns during system design

IEEE P7001/D4 Draft Standard for Transparency of Autonomous Systems		
Summary	Relevant standard in the context of articles 12 "record keeping", 13 "transparency and provision of information to users" and 14 "human oversight" of the AI Act. The aim of this standard is to describe measurable, testable levels of transparency considering the characteristics of system itself as well as those of the key stakeholders involved, enabling the objective assessment of the system through well-defined levels of compliance. The standard supports the principle that it should always be possible to understand "why" and "how" an autonomous system made a particular decision. The standard, however, does not go into details on the technical means to achieve this, implicitly assuming that state-of-the-art techniques exist that can fulfil the defined transparency levels. This may not be obviously the case for some of the strongest requirements defined, which, however, appear to go beyond the demands of the AI Act.	
Artificial Intelligence coverage		The standard focuses on “Autonomous systems” defined as systems that have the capacity to make decisions in response to some input data or stimulus, with a varying degree of human oversight or intervention depending on the system’s level of autonomy. The provided definition, although does not mention specific techniques, should cover the broad range of AI systems considered in the AI Act.
Domain generality		The content of the standard is generally applicable to AI systems across a wide range of sectors. Additionally, it is intended as an "umbrella" standard from which domain-specific standards might develop. It foresees future standards that may be based on this one but cover specific application or technology domains, e.g., standards for transparency in autonomous vehicles, medical or healthcare technologies.
Maturity and level of detail		Mature standard with a considerable level of detail to support transparency, interpretation and explanation of system behaviour by different stakeholders. Among these, most relevant for the AI Act transparency requirements are the different user profiles (lay users, domain experts, technical experts for diagnosis and maintenance). Useful detail is also provided in terms of record keeping, establishing recording requirements that would support system oversight and investigation of incidents. The level of detail for each transparency level and requirement to be achieved is

		reasonable, including specific examples, but it may not be exhaustive, and in some cases they are defined by somewhat ambiguous terms rather than specific metrics or other measurable criteria.
Compliance management		The standard provides well-defined levels of transparency of AI systems to be assessed, which are described in terms of concrete technical functionality. Some of the terms used to describe compliance, such as 'detailed', 'appropriate' or 'accessible', could be more precisely defined in order to ensure objectivity and uniformity in assessing compliance with specific requirements in the AI Act. Notably, it may not be sufficiently obvious how to properly and systematically assess the highest levels of transparency defined, e.g. those that demand sophisticated means like explanations using natural language.
Gaps and complementarities		The document contains well-defined and testable levels of transparency that AI providers can satisfy depending on the intended use of the system. The standard is, however, technology agnostic to a large extent, and does not cover in a meaningful way the technologies needed to support some of the transparency levels. It may be beneficial to complement it with information and guidance for AI providers about transparency and explainability means able to achieve these levels in practice. The informative scenarios covered in the Annex are very relevant (autonomous delivery vehicle, credit scoring system, security robot, medical decision support system, etc.,) but more detail would be beneficial, e.g. describing the systems at the technical level and the transparency measures adopted. Furthermore, this document may need to be complemented with others covering additional aspects of human oversight related to control over the system and the ability of users to intervene in its operation.
Fit within standardisation landscape		Considering the existing landscape of AI standards, this document provides relevant coverage of previously identified gaps regarding transparency, human oversight (with a focus on explainability) and record keeping. This is further detailed in section 5.

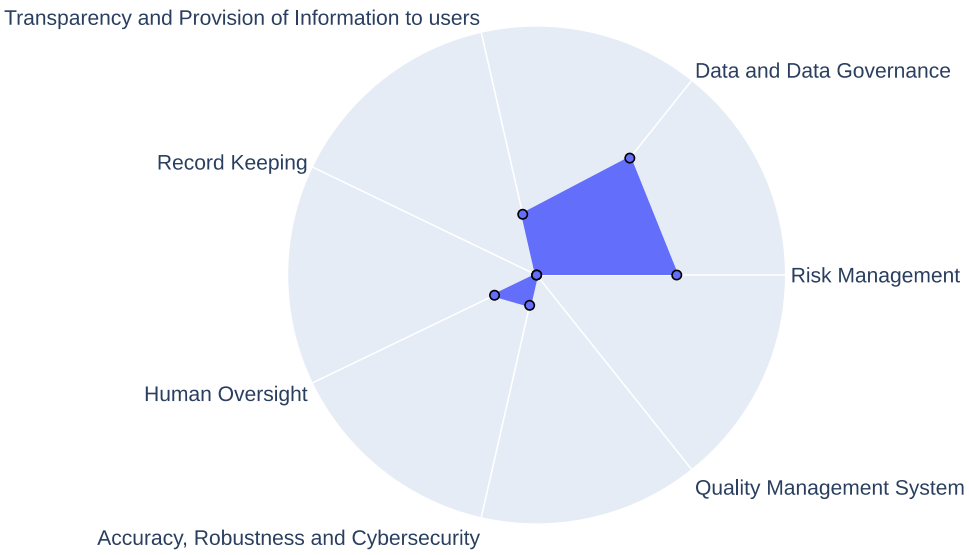
Requirement Coverage



Transparency and Provision of Information to Users	0.22	The standard covers transparency, including interpretation of the system and its intended use by key stakeholders, and explicitly considers the accessibility of users with different profiles (e.g. domain users, technical expert users). Coverage of specific items requested in the AI Act for the instructions of use is partial, as the standard is mostly concerned with interpretation and explanations of AI system behaviour. The specification does, nevertheless, partially cover some relevant aspects related to the characteristics and limitations of the system, such as the intended purpose or the data used for training.
Record Keeping	0.75	Relevant coverage of legal requirements for record keeping and logging, with useful level of detail for AI providers, referring to practical implementation aspects such as the use of standard formats, recording of timestamped data, coverage of inputs, outputs and decisions. Lifecycle coverage is comprehensive and it establishes the need to facilitate inspection and investigation of incidents.
Human Oversight	0.38	The standard covers in detail the needs of users of autonomous systems in terms of information, training material and explanations about system behaviour, enabling them to understand the capabilities and limitations, as well as to monitor the operation of the system. The required human explanation and oversight levels defined go beyond what the AIA demands in practice. However, regarding other relevant aspects, e.g. control and intervention on the operation of the system, coverage seems to be more limited.

Table 4. Analysis of IEEE P7001/D4 - Draft standard for transparency of autonomous systems

IEEE P7003/D1 Draft Standard for Algorithmic Bias Considerations		
Summary	Very relevant standard providing methodologies to identify and address negative bias in algorithmic intelligent systems. It considers different sources of bias, providing suitable courses of actions to detect, assess and mitigate these, such as benchmarking mechanisms, criteria for the creation of validation datasets, guidelines on establishing and communicating application boundaries, and suggestions for expectation management and system interpretation. The framework described is complete in the sense that it addresses not only technical aspects but also organisational ones (e.g. team competencies, accountability mechanisms). It also pays particular attention to the thorough evaluation of biases regarding: (1) stakeholders that might be impacted by (or who may influence) bias in the algorithmic system; and (2) the different operational environments of use (defining testing scenarios accordingly). The standard describes a risk assessment and mitigation plan for bias and considers its documentation at different levels (datasets, algorithms, stakeholders, intended uses, operational settings). It should constitute a valuable reference for AI providers to demonstrate the implementation of best practices in detection and mitigation of negative biases in the design and evaluation of AI systems, especially if complemented with additional guidance regarding the choice of concrete metrics for the evaluation of biases.	
Artificial Intelligence coverage		The standard considers broadly defined "autonomous intelligent systems" which appear to fully cover the definition of AI in the legal text. Proper consideration is given to the risks of machine learning and similar data-driven methods, which are particularly affected by unintended biases. The document also explicitly mentions rules-based systems, statistical systems, and others, while keeping the provisions broadly agnostic to specific algorithmic techniques.
Domain generality		This standard is horizontally applicable to AI systems, independent of the sector in which they operate, and covers all lifecycle stages. It provides a good basis for potential sector-specific standards if needed.
Maturity and level of detail		This standard is still work in progress, and some sections are still not fully developed. It is a process-oriented standard. Some points covered may welcome more detail, e.g. the determination of "justified vs unjustified biases" and "go vs no-go" decisions. On the other hand, some sections appear thorough and complete already, e.g. those covering

		dataset considerations. In general, this standard seems to be on track to provide useful guidance to AI operations in terms of the identification of protected features in data, potential sources and mechanism of bias, their risks to ethical values, and (possibly) on the definition of suitable metrics, e.g. for fairness (still partially underdeveloped). This document should provide relevant content for AI providers to fulfil bias-related requirements in the AI Act spanning several articles, e.g. those dedicated to data and data management, risk management and human oversight.
Compliance management		While still incomplete, the standard contains a well-defined set of activities and tasks to assess compliance, with specific outcomes in terms of documentation and accountability artefacts. Some of these may potentially require more objective criteria, e.g. for assessing the selection of bias and fairness metrics and thresholds.
Gaps and complementarities		Generally complete and thorough coverage of bias considerations in algorithmic systems. However, further detail and concrete examples and best practices may be beneficial for AI providers e.g. regarding the identification of problematic biases in consideration of the intended use of the system and its operational context, the selection of suitable metrics and thresholds, or the creation of balanced datasets. Some of these may be addressed in future drafts of this standard.
Fit within standardisation landscape		This standard addresses a key standardisation gap identified in previous AI standardization analyses, providing extensive coverage of bias considerations in AI systems and complementing and extending existing technical reports on bias in the ISO landscape. This is further detailed in section 5.
Requirement Coverage 		
Risk Management	0.56	The risk management framework described, while limited to bias considerations, is well aligned to the requirements in the legal text, allowing tailoring to the intended use of the system and covering the assessment and mitigation of bias-related risks iteratively. It covers the entire AI lifecycle, including monitoring of performance drift using data collected during operation. Additionally, it provides guidance on the identification of undesired and problematic biases in AI systems and covers testing considerations, e.g. the assessment of datasets as well as algorithm and system outputs, and the definition

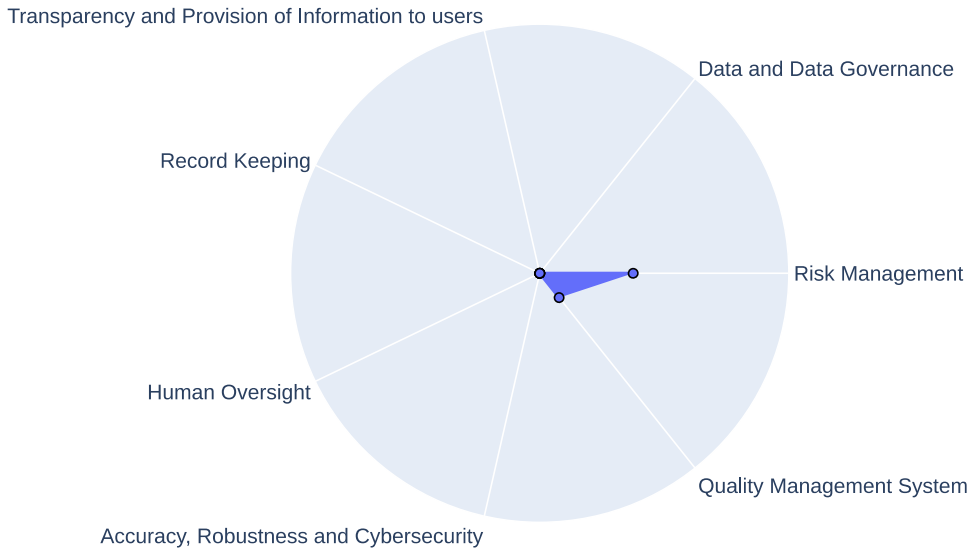
		of test scenarios. It demands definition of bias metrics, including their justification in alignment with the objective use of the system. Finally, it proposes a series of questions to guide the choice of evaluation metrics, and identifies potential automation tools, even if more guidance would be beneficial in this aspect.
Data and Data Governance	0.60	Well aligned with legal requirements in terms of data representativeness and bias. Provides a comprehensive description of bias sources in data, with practical guidance for developers like the consideration of proxies for protected attributes, assessment of data from external sources, consideration of pre-processing in bias mitigations, and general dataset design considerations.
Transparency and Provision of Information to Users	0.25	Contains practical transparency considerations in terms of potential system and data biases as well as details about their evaluation, including the necessary measures of success throughout the lifecycle. The documentation generated from the adoption of this standard is sufficiently detailed for users to understand bias sources, mitigations and limitations, e.g. in terms of residual biases and their justification.
Human Oversight	0.19	This standard covers human oversight requirements related to bias, such as the consideration of complacency bias in system with human oversight, detailing processes for its evaluation.
Accuracy, Robustness and Cybersecurity	0.12	Useful content regarding the selection of suitable accuracy evaluation metrics including full-lifecycle considerations. Considers using operational data to adapt the system behaviour, and the need to periodically review the outcomes so that the system remains within desired performance ranges, preventing performance decay.

Table 5. Analysis of IEEE P7003/D1 - Draft standard for algorithmic bias considerations

IEEE 7010 - Recommended Practice for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-Being		
Summary	The process described in this standard is potentially relevant in the context of risk management requirements, albeit with a narrower focus on measures of human well-being. It defines an impact assessment to produce human well-being indicators, including a broad range of individual and collective well-being considerations, e.g., human rights, economic fairness, social equality, ecological health or access to employment. It describes an iterative process at every phase of the AI lifecycle, from conception, design, development to the ongoing assessment and management of the system. The assessment is data driven, with indicators including subjective as well as objective information collected in operation, e.g. through surveys. However, many of the discussed indicators appear to be suited for rather indirect measures of well-being in statistical terms. Nevertheless, some considerations may be relevant in the context of implementing a post-market monitoring system based on user engagement through focus groups, surveys, experts, crowdsourcing, and similar means. In this context, potentially useful guidance is provided, e.g. considerations about the demographics and representativeness of users and the consideration of potentially underrepresented demographics.	
Artificial Intelligence coverage		The content of the standard is not specific to AI, but it should be broadly applicable to AI systems. The methodology proposed follows relatively generic patterns for evaluation of any kind of application, and the proposed indicators to measure a population's well-being are technology-agnostic.
Domain generality		The process in this document is applicable in a horizontal manner, and it is complemented with informative examples covering specific use cases. The standard describes a range of well-being indicators that could be relevant for different AI systems depending on their intended use and context of operation.
Maturity and level		This process-oriented standard offers a good level of detail, covering well-defined activities and tasks. On the other hand, the indicators provided may be relatively high-

of detail		level when considering the needs of high-risk AI systems in scope of the AI regulation. Some additional guidance on the criteria to select appropriate indicators (or how to define new relevant ones) for these risks may be needed. Some of the examples provided (facial recognition, hiring, healthcare) are relevant in the context of the high-risk AI systems defined in the regulation.
Compliance management		Determining compliance with the method and steps defined in this standard should be reasonably uncomplicated. However, assessment of the suitability of the well-being criteria selected for a given AI system is out of scope, and these criteria may be strongly linked to adherence with legal requirements.
Gaps and complementarities		Well-being indicators are described in a thorough manner. However, many of them appear to be tailored to measure the impact of AI in a statistical sense, by user reporting of perceived levels of satisfaction and well-being after system adoption. In this way, the standard is extremely generic and able to assess any system, whether AI based or not. However, it appears to be mostly suitable to capture trends derived from AI adoption at scale rather than the risks and impact of individual products. Therefore, this standard would need to be complemented by more direct risk-assessment methods. Nevertheless, the methodology described may play a role in some AI systems where direct assessment of risks is not possible or may be incomplete. In these cases, the open-ended nature of the well-being questions defined in this standard may be of help in identifying otherwise missed negative impacts of AI.
Fit within standardisation landscape		This standard provides potentially relevant human well-being indicators not covered by existing specifications, and parts of it may be useful to inform certain post-market monitoring activities defined in the AI Act.

Requirement Coverage



Risk Management	0.38	This standard offers some potentially relevant processes for risk management that emphasise the iterative and continual evaluation of the risks and impact of a system on human well-being. This methodology should be suited to assess possible negative AI impacts in operation, albeit in a statistical sense. This process should be suited to

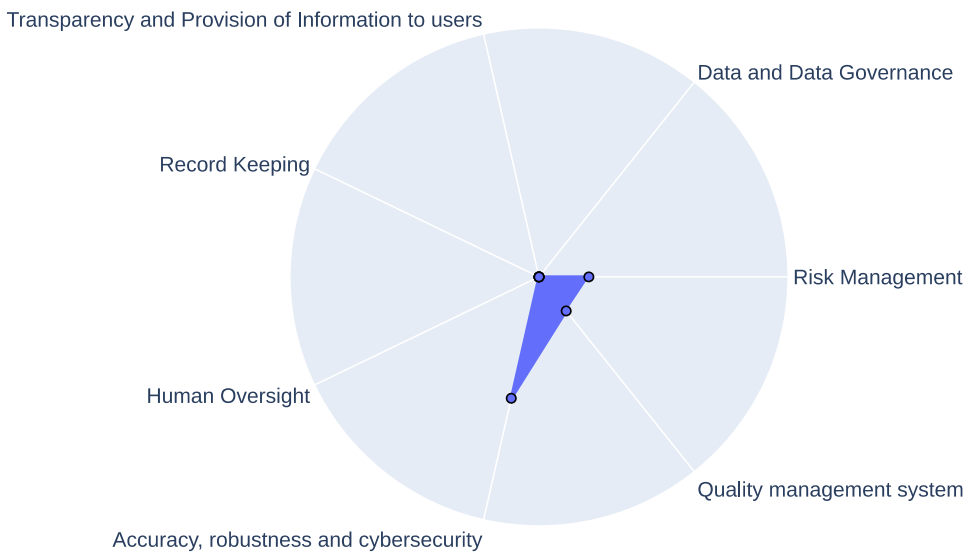
		uncover negative impacts resulting from changes in the context of use, or those that may occur in AI systems that continuously learn and adapt based on interaction with users. It acknowledges that the selection of indicators for well-being that an AI provider should use depend on the nature of the system in question and the circumstances of those potentially impacted. The described impact assessment method and its implementation provide useful guidance, including for the selection of well-being indicators as well as data collection and analysis to improve the system.
Quality Management System	0.12	Some clauses of these standard could be relevant to inform the setup, implementation and maintenance of a post-market monitoring system, albeit with a narrow focus on high-level human well-being indicators. It considers gathering information for system monitoring and improvement. However, the indicators described seem rather generic, and may be best used to statistically measure the impact of systems adopted as scale, rather than being able to identify concrete incidents and failures in individual AI systems.

Table 6. Analysis of IEEE 7010 - Recommended practice for assessing the impact of autonomous and intelligent systems on human well-being

IEEE P2841 - Framework and Process for Deep Learning Evaluation		
Summary	The objective of the IEEE 2841 standard is to support providers of systems employing deep learning to assess the reliability of their algorithms with a set of processes and indexes, with the objective of improving the resulting software quality. The standard lists concrete activities that AI providers can implement covering different development stages from design to implementation and operational aspects, hence covering the entire AI system lifecycle. It is largely organizational in nature, proposing a system of indexes assessing the correctness of algorithm function implementation, correctness of code and the influence of different design elements and environmental influences such as objective functions, training data, hardware, adversarial examples, software platforms or environmental data. These could guide developers to conduct reliability assessment and hence this standard is relevant in the context of Chapter 2 requirements in the AI Act, with a focus on Article 15, especially the robustness and cybersecurity provisions, given the document's focus on concrete failure modes. In this sense, the standard appears to favour breadth over technical depth, resulting in the relevant technical elements being presented at a high level. Therefore, this document may benefit from additional detail and technical guidance, in order to effectively guide AI developers to implement the processes described in practice.	
Artificial Intelligence coverage		The aim of the standard is to focus solely on deep learning systems, but, in fact, the described assessment indices and processes seem broadly applicable to other machine learning systems and even other types of AI algorithmic systems for classification. However, this comes at the expense of not addressing many specificities of deep learning systems. Many of the considerations of the standard seem to explicitly focus on supervised machine learning classification methods, thus leaving out of scope other approaches in machine learning such as regression problems, unsupervised and reinforcement learning.
Domain generality		The standard is horizontally applicable to components of high risk AI systems using machine learning independent of the application domain.
Maturity and level of detail		The described processes, methodologies and organizational principles behind the suggested assessment process of machine learning models is mostly clear and moderately detailed. It covers therefore the AI lifecycle comprehensively. Especially the risk and hazard assessment process laid down in Section 5 and the detailed matrix in Appendix A for the choice of assessment indices according to process stage and risk could be of practical use as they are. The list of assessment indices appears to be broadly complete and not purely limited to deep learning considerations.

		General concerns covered include correctness of code, code and AI vulnerabilities and practical AI considerations such as dependencies introduced by deep learning frameworks, hardware architectures and devices or operating systems. These are, however, covered rather superficially, and more detail would be beneficial to ensure that adoption of the standard leads to a consistent and effective practical implementation. In this sense, guidance on state of the art techniques, possible technical procedures or metrics is scarce. This may not be critical for certain technical parts of the assessment checklist, where the application itself typically demands specific choices, e.g. performance metrics selection or the analysis of target function effects with regards to overfitting. In other cases, the lack of technical detail may need to be filled by other standards in order to fulfil the requirements of the AIA as no common or standardized practice have been established, e.g. in terms of cybersecurity assessment, selection of reliability targets, datasets and target functions assessment or evaluation against adversarial examples.
Compliance management		Demonstrating compliance with the organizational principles in this standard should be straightforward given the well-defined assessment processes described in sections 5-9 and Appendix A. It could also easily and transparently be expanded, by adding more assessment indices. On the other hand, the superficial, non-prescriptive definition of the assessment indices may make it difficult to ensure consistent adoption and hence to guarantee that AI systems assessed on the basis of this standard are uniformly and consistently robust.
Gaps and Complementarities		The standard focuses on a reliability assessment for a single deep learning model, whereas it is conceivable that most practical high-risk AI applications will be contain multiple software and IT infrastructure components, hence demanding a system-level assessment procedure to complement the model-centric evaluation proposed by this standard. Additionally, some concrete Machine Learning approaches such as regression problems, online learning systems, unsupervised learning or reinforcement learning may deserve more consideration by the standard, which appears to focus on supervised classification approaches.
Fit within standardisation landscape		Assessment of AI (including deep learning) systems is already addressed by existing international standards from ISO/IEC, which, besides providing broader coverage of AI techniques, contain technical additional detail not present in the IEEE standard. This includes, for example, the ISO/IEC 24029 documents on robustness aspects, as well as the ISO/IEC 4213.2 on Machine Learning classification performance, which are more technically oriented. Some of these ISO standards appear to go beyond the IEEE document in terms of guiding the selection of accuracy and robustness metrics. Other ISO technical reports on AI system software testing such as ISO/IEC 29119-11 also present a substantial overlap with the IEEE standard, covering similar topics as the core part of IEEE-P2841. Despite this, the IEEE document contains complementary material, e.g. the assessment index concept, and the related assessment table for different hazard severity scenarios in the different stages of the AI lifecycle, which could be useful to formalize AI system assessment in practical terms. However, concrete processes and guidelines on testing of AI may also be shortly provided by upcoming ISO standards on AI assessment, e.g. a new work item proposal ISO/IEC NP TS 17847 on verification and validation analysis of AI systems.

Requirement Coverage

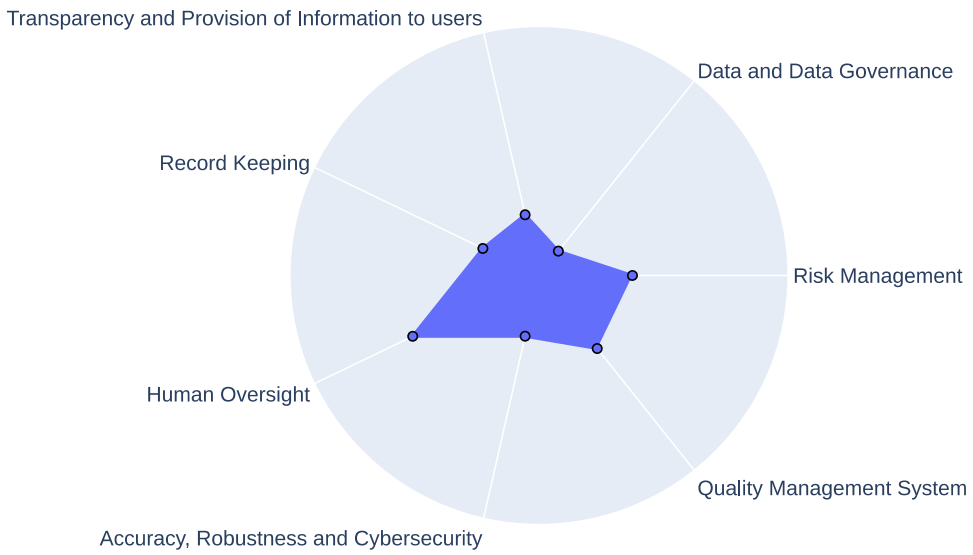


Risk Management	0.20	Clauses 5-8 of this document cover relevant risk management aspects, from an organizational perspective, that could help AI providers identify foreseeable risks in high-risk AI systems making use of deep learning. Test-related requirements in Article 9 of the AI Act are also partially covered in this document, including the definition of suitable testing procedures according to the intended purpose and risks posed by the AI system, even if technical detail is generally scarce. An important element which is not sufficiently considered in this standard is the iterative nature of risk management throughout the AI lifecycle stages.
Accuracy, robustness and cybersecurity	0.50	The standard provides, mostly in section 4, broad coverage of the requirements in Article 15, including accuracy, robustness and cybersecurity. The standard would benefit from more technical depth, e.g. by describing practical methods for the verification and empirical testing of large deep learning models, including analysis of robustness, adversarial examples, data poisoning or classical security vulnerabilities. While providing solid organisational and procedural measures, the standard does not specify the technical solutions, or these are covered only at a high level. Some of the requirements of the AI Act may be missing, e.g. guidance on the selection of relevant accuracy metrics, or robustness approaches for systems using online/continuous learning or requiring redundancy or fail-safe mechanisms.
Quality management system	0.17	The assessment process described, specifically in sections 4-9, could fulfil some of the requirements in Article 17 of the AI act towards implementing a quality management process, especially regarding accuracy, robustness and cybersecurity of AI systems. This includes the application and documentation of specific techniques, procedures and actions during the design, development and verification of high-risk AI systems. Other aspects, notably data-related aspects, appear to be only minimally considered by the document, limited to certain considerations for training and test data handling.

Table 7. Analysis of IEEE P2841 - Framework and Process for Deep Learning Evaluation

IEEE ECPAIS - Accountability Certification Requirements		
Summary	This certification suite describes criteria to assess the accountability of organisations dealing with Autonomous and Intelligent Systems, which includes AI systems. This document lists requirements and evidence generally applicable to most settings involving AI systems, demanding a clear distribution of responsibilities and tasks among the organisations' staff to ensure accountability. It identifies 8 top-level aspects related to accountability, including 5 drivers (organisational governance, clarity of operations, human oversight, user interactions, upholding ethical profile) and three inhibitors (random and systematic errors, rubber stamping and inadequate or non-existent records). A subset of these criteria is relevant in the context of various requirements for high-risk AI systems defined in the AI Act and could be useful in the context of the specific responsibility of providers of high-risk AI systems to set up a quality management system including an accountability framework. As outlined below, for some of the certification criteria defined it would be beneficial to provide more concrete details and guidance for the AI providers and assessors expected to produce and evaluate the evidence listed. Some of these are expected to be provided in additional AIS certification resources developed and made available by IEEE after this review was carried out.	
Artificial Intelligence coverage		These criteria are specific to autonomous and intelligent systems, which encompass AI systems. This certification suite should apply to all AI systems independent of risk levels, and some of them are specific for higher-risk AI systems.
Domain generality		The certification suite contains top-level, non-sector specific requirements and evidence that can be horizontally applied to a wide range of AI applications.
Maturity and level of detail		While many of the specified criteria contain some useful elements, further selection, specification and tailoring would be beneficial in order to turn this list of criteria into a suitable and operational means for assessing compliance with legal requirements defined in the AI Act. It should be noted that a full implementation process for ECPAIS criteria has been developed by IEEE which could address this point. This is discussed in section 6. Approximately half of the criteria was perceived not to be directly applicable to requirements for high-risk AI systems defined in the AI Act requirements or were considered repetitive, abstract or unspecific. Criteria G2 (Clarity of Operations), G3 (Human Oversight) and parts of G5 (Upholding Ethical Profile) and G1 (Random and Systematic Errors) concentrate most of the relevant elements. The rationale for roles assigned to criteria is not clear, and some further information would be beneficial regarding specific responsibilities for the different requirements and evidence requested.
Compliance management		The evidence required for assessment is in line with documentation and quality management requirements in the legal text, even if some of the evidence goes beyond, e.g. including meeting minutes, audit reports, external consultant studies, interviews with staff. The measurement scale defined to evaluate evidence appears at this stage subjective and up to individual evaluators, and is expected to be adapted and fine-tuned in the future based on experience. The items listed as "acceptable evidence" appear to be feasible to verify. Especially, third-level criteria, which are to a good part instructive instead of normative, appear to have more detailed content. However, many elements could be further improved by defining measurable and objective evidence.
Gaps and complementarities		This document could contribute useful criteria towards setting up and assessing an accountability framework covering AI Act-defined requirements, e.g. related to risk management, transparency, testing and human oversight. One aspect found not to be sufficiently covered in the criteria is that of data management and governance, e.g. responsibilities related to data collection, dataset design and curation or quality assessment.

Fit within standardisation landscape		Review of this document has shown some degree of coverage of all of the requirements for high-risk AI systems in the AI Act. Given its focus on breadth, the document does not provide substantial in-depth coverage of any particular requirements. Despite this, it could be a useful source to complement existing standards covering organizational aspects of AI, and as outlined in this detailed report, and some of the criteria could be selected as part of a certification suite tailored to the specific needs of the AI Act.
Requirement Coverage		
Risk Management	0.38	Some elements are relevant for risk management, e.g. part of G2 (Clarity of Operations), which could inform the set-up and assessment of a risk management system, including details on the proper design of a testing protocol, or parts of G5 (Upholding Ethical Profile) related to risk management and risk assessment. The evidence described may be useful to enable uniform conformity assessment, even if the process for providers to generate it is not fully detailed as part of the criteria.
Data and Data Governance	0.12	Some elements, e.g. from G1 (Random and Systematic Errors), may be relevant for verifying accountability with regards to data management, e.g. providing documented evidence about the achievement of fairness and bias objectives, providing records of datasets, or evidence of changes to address errors. The criteria include some useful, even if superficial, guidance on how to identify steps to be taken, e.g. data shuffling, addressing sampling errors, addressing bias and fairness, address poor data quality, documenting data and measuring the effectiveness of data enhancements.
Transparency and Provision of Information to Users	0.25	Criteria defined in G2 (Clarity of Operation) are relevant, e.g. covering the different modes of intended operation, operational environments, stakeholders and contexts analysed under various possible scenarios. Some of the evidence described would cover needs defined in the AIA for instructions of use, e.g. related to intended purpose, performance or potential risks from changes after market placement.
Record Keeping	0.25	Certain elements, also defined mostly from the G2 (Clarity of Operation) list of criteria, contain useful elements to fulfil and verify certain record-keeping



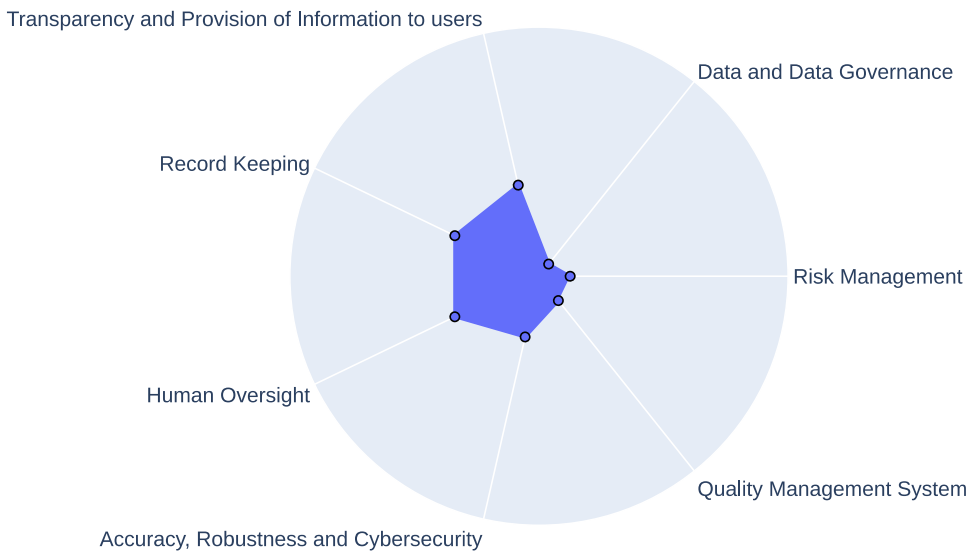
		requirements, detailing potentially useful means and evidence, e.g. the recording of system inputs, outputs, errors, issues, malfunctions and modifications to the AI system. It envisions potentially relevant means to present information, e.g. in the form of dashboards for the monitoring of potential deviations.
Human Oversight	0.56	A subset of the assessment criteria in G2 (Clarity of Operation) and G3 (Human Oversight), contain relevant human oversight elements, e.g. through evidence of suitable human involvement in the operation of the system, ensuring a proper understanding of the specific decisions that are automated, and their limits. This includes means for not overly relying on the AI, i.e. promoting human discretion and judgement with an active rather than passive role, and awareness of clear lines of responsibility. Other relevant evidence described for AI providers includes the definition of thresholds in systems to ensure human intervention when needed, e.g. through mechanisms that provide an early warning to operators or otherwise support in the identification of relevant deviations to identify appropriate action.
Accuracy, Robustness and Cybersecurity	0.25	Partially relevant considerations in G2 (Clarity of Operation) that may support and provide evidence of accuracy and resilience objectives, even if described at a high level. Relevant requirements and evidence include, e.g. testing against operational context prior to deployment, cross verification against specifications, periodic testing and validation evidence, parameter tuning documentation and definition of modification thresholds, including in systems that continue to learn after being placed in operation. It also provides coverage of evidence related to cybersecurity requirements, e.g. records of the tracking and pre-emption of adversarial examples, and protections against the alteration of datasets and algorithms.
Quality Management System	0.38	There are several useful elements throughout this certification criteria suite that could enable providers of high-risk AI systems set up and document an accountability framework adapted to the needs specified in Article 17 of the AIA Act, covering e.g. risk management aspects, post-market monitoring considerations, reporting and tracking of incidents and malfunctions or certain data management elements.

Table 8. Analysis of IEEE ECPAIS Accountability Certification Requirements

IEEE ECPAIS - Transparency Certification Requirements		
Summary	This document provides a list of criteria to certify autonomous and intelligence systems for transparency. It identifies 9 top-level aspects that can be used by certifiers to predict transparency in AI systems, including 6 drivers (organisational governance, clarity of operations, awareness of A/IS interaction, confidence in system behaviour, accessible control & feedback and upholding ethical integrity) and 3 inhibitors (behavioural obfuscation, concern with liability and protection of trade secrets). Many of these contain elements that are not directly related to the requirements in the AI Act. However, a small subset of the criteria (e.g. related to clarity of operations and confidence in system behaviour) is relevant to the AI Act. These could be part of a more concise assessment suite for the operationalization of some requirements for high-risk AI systems, most notably those related to human oversight, transparency and provision of information to users and record keeping. However, this would require more concreteness in the definition of the criteria to assess and score the evidence described in this document. Additional material made available by IEEE after this review, defining a process for the tailoring and application of the certification criteria in practice, could be useful in this regard.	
Artificial Intelligence coverage		This certification criteria covers autonomous and intelligent systems, and most considerations should apply to AI systems horizontally (non-sector specific). A subset of the criteria covers specific AI techniques (e.g. federated learning).
Domain generality		The certification criteria are in principle applicable to high-risk AI systems operating in all of the sectors and domains considered in the AI regulation.

Maturity and level of detail		A small subset of the criteria, under G2 (clarity of operations) and G4 (confidence in system behaviour), were found to be relevant to the requirements in the AI Act. Overall, the document contains some repetitive elements and may be partially out of date in terms of content and terminology. Some important terms related to transparency, e.g. those concerning the explainability and interpretability of AI models, are not considered or sufficiently detailed. To a large extent, this is reasonable and expected given the completion date of the document (2019) and the rapid advancement of AI transparency and human oversight techniques, and may indicate the need to update these criteria to reflect advances in the state of the art.
Compliance management		The document describes concrete requirements that the AI system must fulfil and the evidence required to demonstrate compliance. The forms of evidence appear to be in line with documentation obligations in the legal text. However, more detail in the content to be demanded would be beneficial, including how to assess the described evidence at the technical level. In many cases this is specified in an abstract manner, using terms such as "appropriate", "adequate", "employ a mechanism", "desirable performance", "be mindful of", "right mechanism" or "well communicated". Furthermore, at this stage there are limited indications on how to assign scores, so the assessment could depend to a large extent on individual assessors. It should be noted that a full implementation process for ECPAIS criteria has been developed by IEEE which could address some of these points. This is discussed in section 6.
Gaps and complementarities		While some of the sections are relevant and potentially useful, the document appears to go for breadth, covering many aspects at once. As a consequence, in terms of transparency and human oversight requirements defined in the AI Act, the level of depth and coverage of the state of the art could in many cases be improved.
Fit within standardisation landscape		Similar to other certification criteria suites reviewed, the content of this document spans many aspects of AI systems. Some of these, especially those related to human oversight and record keeping may contribute towards filling known standardisation gaps, although greater depth and technical specificity would be beneficial.

Requirement Coverage



Risk Management	0.12	Some of the organisational criteria are potentially relevant in the context of providing evidence about transparency risks and associated mitigations.
Data and Data Governance	0.06	Some relevant even if superficial coverage of some data and data governance requirements, e.g. evidence of completeness and accuracy of datasets, or consideration of geographical concerns.
Transparency and Provision of Information to Users	0.38	Assessment criteria under G2 (clarity of operations) and G4 (confidence in system behaviour) offer potentially useful details towards providing accessible information to users, covering among others, system capabilities and performance, behavioural features, concept of operation, functional design, manuals and guidelines. It covers certain relevant aspects for users of high-risk AI systems as required in article 13 of the AI Act.
Record Keeping	0.38	Criteria under G2 (clarify of operations) and G4 (confidence in system behaviour) also contain relevant considerations and evidence related to record keeping, e.g. sampling and recording input and output data or the provision of immutable records. In this aspect it goes beyond listing desired capabilities and provides some useful coverage of technological measures.
Human Oversight	0.38	Assessment criteria covering human oversight aspects can be found under G2 (clarity of operations) and G3 (awareness of A/IS interaction), with potentially relevant considerations and technical measures, e.g. the provision of system information via APIs and dashboards, the necessity to monitor performance and robustness on the user side or to account for potential false positives and alarms, and the implementation of mechanisms to identify safety concerns and trigger system inspections.
Accuracy, Robustness and Cybersecurity	0.25	Some of the considerations under G2 (clarify of operations) and G4 (confidence in system behaviour) appear relevant to the evaluation of system behaviour. Evidence such as test reports are in line with the documentation demanded by the AI Act. Some potentially useful guidance for providers is contained, e.g. regarding target levels of statistical significance for tests. Evidence of infrastructure hardening measures, relevant to cybersecurity requirements, are also part of this criteria, although the objectives are described in rather abstract terms.
Quality Management System	0.12	The organizational criteria under G1 contains a small set of potentially relevant considerations for the setup of an accountability framework in line with the quality management requirements in the legal text, with a focus on transparency aspects, even if further detail on the assessment of the described evidence would be beneficial.

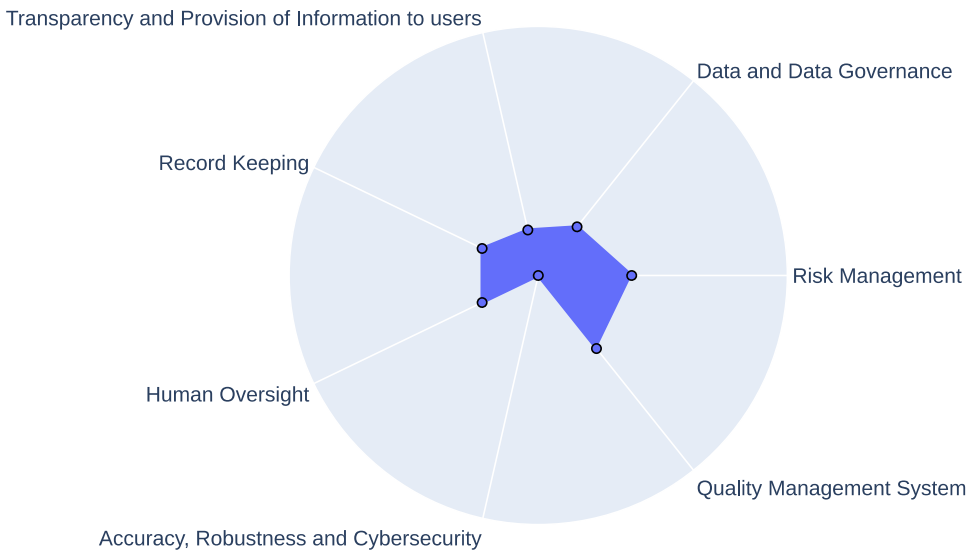
Table 9. Analysis of IEEE ECPAIS Transparency Certification Requirements

IEEE ECPAIS - Bias Certification Requirements	
Summary	This set of certification criteria defines measurable predictors for assessing bias in autonomous and intelligent systems, including 6 drivers (suitable organisational governance, clarity of concept and operation, appropriate context alignment, justified protected characteristics, system behaviour monitoring and maintaining an acceptable bias profile) and one inhibitor (lack of process transparency). The criteria defined is broadly relevant to the AI Act, touching on various legal requirements for high-risk AI systems from the perspective of bias. The criteria and associated evidence described are broad ranging, e.g. including policies, procedures and instructions for managing bias-related risks, design controls and verification aspects to mitigate bias, considerations on stakeholders' ability to identify and implement bias mitigations, measures to facilitate transparency and understanding of the system's context of use and its relevant operational environments, provisions to enable the adaptation of systems to new operation contexts, measures related to the establishment of appropriate feedback channels with stakeholders, considerations related to the use of protected characteristics in algorithms,

	and system monitoring requirements including information recording and human intervention considerations. Some areas where further elaboration would be useful include coverage of data-related considerations for bias (e.g. in datasets used for training models) and the selection of concrete metrics to measure and monitor bias.	
Artificial Intelligence coverage		In line with other IEEE documents, this certification criteria applies to "Autonomous and/or Intelligent Systems" with a definition that broadly covers AI systems considered in the AI Act.
Domain generality		This document can be applied to AI systems horizontally. Criteria are multi-domain, and indeed an important part of the standard aims at assessing (1) contexts of use, (2) impacted stakeholders (identifying their needs and expectations) and (3) operational environments. The definition of bias redlines, intervention triggers and justified biases shall be made according (and in adequate proportionality) to these three points.
Maturity and level of detail		The document provides a comprehensive list of assessment criteria for bias in AI systems. The organizational elements listed include measures to identify stakeholders, assess AI system impact, and ensure transparency and compliance with the legal environment, as well as measures to ensure that development teams have adequate training and skills to understand and manage biases. Transparency considerations appear to be thorough, even if at times overly generic. The section dedicated to ensuring appropriate context alignment offers relevant detail for AI providers, e.g. in terms of evidence to provide regarding identified biases, their translation of ethical requirements, the consideration of local tuning aspects and the creation of testing protocols that take into account cultural, social, geographical and legal differences between stakeholders. On the other hand, sections covering more technical and direct measures of bias offer less detail, for example, the use of protected attributes is only briefly covered. Similarly, even though bias monitoring and testing reasonably spans the AI system lifecycle, provisions on dataset aspects is superficial and could be considerably expanded. The content related to bias profiles could be useful to inform and assess the management of bias-related risks and to ensure the right processes to correct emerging or detected bias during development, deployment and operation are in place.
Compliance management		The document defines processes and methodologies to assess AI systems against bias requirements. Many of them contain useful detail on practical means to operationalize these requirements and to provide evidence that biases are understood, monitored and addressed through the design and operation of AI systems. On the other hand, the description of acceptable evidence could be further detailed, with concrete guidance for assessors to evaluate them in an objective and effective manner, i.e. for the identification of concrete predictors for bias in the documentation and test artefacts described. These could include more direct and objective bias measurement and assessment factors, e.g. how to select and assess bias metrics, thresholds and triggers for interventions, or suitable evaluation results (ROC curves, confusion matrices, F1 scores, data distributions, etc.) for bias. While some of these may be included in subsequent, sector-specific documents, some degree of horizontal coverage of such technical assessment means for bias could increase the effectiveness of this certification suite.
Gaps and complementarities		The document covers the entire AI system lifecycle reasonably well, providing a good trade-off between pre- and post-deployment criteria. The balance between organizational/governance-related and technical criteria seems to be in favour of the former and could therefore perhaps be improved for the latter. Considering the requirements for high-risk AI systems in the AI Act, the bias assessment criteria could be complemented with additional considerations for data and data governance, including technical means to assess, monitor and mitigate bias in datasets used to train and validate AI systems.

Fit within standardisation landscape		This document provides criteria for assessing bias in AI systems, a prominent consideration in the AI regulation in need of additional technical specifications.

Requirement Coverage



Risk Management	0.38	Risk management considerations in the AI Act are well covered with respect to bias risks in this certification suite. Relevant criteria are found across different sections. A significant driver is G6 (acceptable bias profile), which can inform the assessment of risk management measures for bias, e.g. processes to correct emerging or detected bias during development, deployment and operation, the implementation of algorithmic impact assessments and similar ongoing review of system, identifying relevant affected stakeholders. Other relevant sections include G5 (system behaviour monitoring) for testing-related considerations for bias, and G3 (appropriate context alignment) defining relevant evidence for the consideration and mitigation of bias, including inventories of identified biases, ethical requirements derived from them or evidence of local tuning. Finally, G1 (organisational governance) as well as G2 (clarity of concept of operation) may provide useful considerations related to organizational aspects for the identification and transparency of bias-related risks.
Data and Data Governance	0.25	Data and data governance aspects appear to be relatively underrepresented for an assessment suite on bias. Mostly G5 (system behaviour monitoring) and parts of G3 (appropriate context alignment) and G6 (maintaining an acceptable bias profile) have some provisions on datasets used for training, validation and testing, including monitoring for consistency with data from operation, and assessment of bias impact of new collected data. Coverage, however, is arguably low, especially considering the potential impact of training datasets in AI system bias.
Transparency and Provision of Information to Users	0.19	Relevant criteria defined in G2 (clarity of concept of operation) cover transparency and technical documentation of technical aspects. Some are specific to bias but several appear to be rather generic documentation elements about the system, tests, audits, and context of use.

Record Keeping	0.25	Some potentially relevant provisions in terms of assessing logging capabilities and traceability of the system are included in G5 (system behaviour monitoring), with a focus on the identification of biases and measures such as intervention triggers.
Human Oversight	0.25	Criteria defined in G5 (system behaviour monitoring) provides some coverage of human oversight with respect to biases, considering measures for the identification, monitoring and redress of biased behaviours.
Quality Management System	0.38	The evidence included in G5 (system behaviour monitoring) and especially G6 (maintaining an acceptable bias profile) is relevant in the context of the quality management system defined in the AI Act, e.g. including evidence of monitoring for bias during the entire lifecycle, including development, deployment and post-market operation phases, as well as for the reporting of incidents related to bias. Some of the criteria defined in G1 (Suitable & Sufficient Organizational Governance) could also serve as a basis to define evidence related to accountability for bias, e.g. terms of policies, procedures and instructions for managing bias-related risks, including design, control and verification aspects to mitigate bias.

Table 10. Analysis of IEEE ECPAIS Bias Certification Requirements

5 Standards for the European AI Act: IEEE and ISO/IEC complementarities

Previous analysis already identifies certain areas where IEEE standards provide relevant coverage of AI Act requirements, along with relevant ISO/IEC standards in their respective technical areas. Some of these standards appear to complement the ISO/IEC JTC1 SC42 landscape particularly well, representing valuable sources to be considered for adoption in the European context by the European Standardisation Organizations. In particular, the following 3 standards are considered very relevant:

- **IEEE P7003 "Standard for Algorithmic Bias Considerations"**. Assessment and mitigation of unwanted bias is one of the most fundamental concerns in the design of AI systems and, while not being the subject of a dedicated article in the legal text, it significantly impacts all the trustworthiness requirements for high-risk AI systems. While still in draft form, IEEE P7003 appears to be on course to provide a comprehensive coverage of bias in algorithmic systems and should be considered a priority reference for bias-related requirements in the future AI Act. Subsection 5.1 provides a comparison of this standard with related ISO/IEC work, including the ISO/IEC TR 24027 technical report on "Bias in AI systems and AI aided decision making" and ISO/IEC DTS 12791 "Treatment of unwanted bias in classification and regression machine learning tasks", a technical specification in an early development stage,

- **IEEE P7001/D4 "Draft Standard for Transparency of Autonomous Systems".** This standard provides partial coverage of requirements on transparency, record-keeping and human oversight.

  Regarding transparency, the IEEE P7001 standard provides partial coverage of the requirements in Article 13 of the AI Act. It considers users of AI systems as relevant stakeholders, and defines relevant transparency means such as instructions for use. However, the coverage of specific information items requested in the AI Act for the instructions of use is incomplete. Full coverage of transparency requirements may have to be filled either directly through European Standardisation, or by upcoming ISO/IEC work, such as ISO/IEC AWI 12792 "Transparency taxonomy of AI systems". In this context, the formalization of industry practices for AI system documentation may be beneficial, as several of them have been found to be relevant in the context of the AI regulation [11].

  When it comes to record keeping, its coverage in IEEE P7001 is significant. While there may be existing standards covering general concerns, e.g. logging formats or record keeping approaches applicable to most software systems, there is, in the standardisation landscape reviewed so far, an apparent lack of standards providing AI-specific record keeping technical specification. These include, for example, specific guidance on recording AI system decisions, intermediate states and internal events in a comprehensive manner. Coverage of some of these elements makes IEEE P7001 a particularly relevant source to consider.

  Finally, human oversight is also an area covered in IEEE P7001. In this context, the provisions in this standard related to enabling understanding of the system's function by its users and operators, are directly relevant in the context of Article 14 of the AI Act. In addition, considerations related to explaining the system's decisions could also play a role. A more detailed discussion on the relevance of explainability techniques such as the ones presented in IEEE P7001 as well as ISO/IEC TS 6254 "Objectives and approaches for explainability of ML models and AI systems" is provided in subsection 5.2. Independent of this, it should be noted that other aspects of human oversight, such as controllability and related approaches to intervene on the operation of AI systems, do not appear to be covered in a meaningful way in these specifications, and may require consideration of future standards, e.g. ISO/IEC AWI TS 8200 "Controllability of automated artificial intelligence systems".

- **IEEE 7000 "Standard Model Process for Addressing Ethical Concerns during System Design".** This standard is fundamentally about ethical design and has been found especially relevant in the context of risk management requirements in Article 9 of the AI regulation. In the ISO/IEC context we can find similarly mature and relevant documents such as ISO/IEC 23894 "Artificial Intelligence - Risk Management" as well as parts of ISO/IEC CD 42001 "Artificial Intelligence – Management system". However, the product design-oriented nature of IEEE 7000 provides a unique point of view that may render it especially useful for providers of high-risk AI systems looking to integrate risk considerations into the early stages of their design lifecycle. Furthermore, IEEE 7000 is currently undergoing the adoption process at the ISO/IEC level. This makes this standard a highly relevant one to consider for future European adoption in the context of the AI Act, provided that the limitations identified are addressed in the process. A further analysis of complementarities between IEEE 7000 and ISO/IEC 23894 can be found in subsection 5.3.

5.1 Standardisation of bias assessment and mitigation

Addressing potentially harmful effects of bias in AI is of key importance for the development of trustworthy AI. In fact, the very use of the term *bias* in the AI Act is closely linked to risks to fundamental rights, such as discrimination and other unfair differences in treatment. In this regard, it should be noted that standards tend to differentiate between a technical and neutral definition of bias, something necessary and inherent to AI systems, and either "unwanted" bias (ISO/IEC) or "unjustified" and "inappropriate" bias (IEEE), more closely linked to the use of the term in the legal text. All the standards reviewed identify these unwanted forms of bias as the target of assessment and mitigation approaches, including both technical and non-technical measures, in order to prevent negative outcomes derived from the use of AI systems.

In light of this, bias assessment and mitigation methods and techniques are expected to appear prominently in European and harmonized standards for the AI Act, including preliminary work items currently in the roadmap of European Standardisation Organizations, such as CEN-CENELEC-ETSI PWI "Overarching unified approach on trustworthiness characteristics". However, European standardisation work on AI bias does not need to start from scratch, as our analysis of the IEEE standard in this area shows. Indeed, IEEE 7003, while not yet completed, is on course to provide a thorough coverage of bias identification, assessment and mitigation in algorithmic systems, representing a useful resource for AI providers to prevent their systems introducing unintended, unjustified or unacceptable biases in decision making. Furthermore, this specification is well matched with existing and ongoing work on AI bias at the ISO/IEC level. An example is ISO/IEC DTR 24027 "Bias in AI systems and AI aided decision making", which, while not being a prescriptive document, offers a comprehensive description of sources of unwanted biases in AI, including concrete methods and metrics for assessing and treating them, a concrete aspect where IEEE work could be complemented. Further relevant material is expected to be developed in the context of ISO/IEC 12791 "Treatment of unwanted bias in classification and regression machine learning tasks", a recently started technical specification building on ISO/IEC 24027 and focusing on machine learning. At this stage, ISO/IEC 12791 contains little more than an outline, but it aims to provide standard mitigation techniques to be applied throughout the entire AI system life cycle to treat unwanted bias, addressing stages after deployment, such as operations and post-market monitoring and even retirement of the AI system, as demanded by the AI Act.

In addition, complementarities between IEEE and ISO/IEC work on bias may extend beyond the substance of these standards, extending to their structure for presenting bias considerations to stakeholders. In this sense, ISO/IEC work categorizes sources of unwanted bias according to where they originate in the AI system: be it individuals, data or engineering processes, and can be naturally used to identify the most relevant sources of bias depending on the specific AI lifecycle stage. On the other hand, IEEE focuses on the mechanisms that result in unwanted bias, such as omission or skew, as classifying criteria. IEEE 7003 is particularly useful as a comprehensive catalogue and checklist to understand, assess and mitigate the different sources of bias. Both ISO and IEEE perspectives, respectively centred on the AI-lifecycle and the mechanisms of bias, are valuable for providers of high-risk AI systems. They both agree on the fundamentals, such as the fact that bias depends on societal factors as well as on computational ones, and both stress the need for methodologies to be tuned to specific cases and contexts of use of AI systems.

In conclusion, IEEE 7003 should be considered a very relevant source for standardisation of AI Act requirements in relation to bias, providing a broad coverage of the topic as outlined in the review presented in section 4. Furthermore, this standard can potentially be complemented by upcoming ISO/IEC documents, for example in terms of providing further guidance for the identification of bias in concrete steps of the AI lifecycle.

5.2 AI transparency and the role of explainability standards

Articles 13 and 14 of the AI Act prescribe that users of high-risk AI systems should be able to interpret its output with the objective of ensuring appropriate use and human oversight. Indeed, transparency of an AI system is instrumental for users to understand and use its outputs appropriately, as well as to oversee its operation. To this end, the field of explainable AI has produced relevant techniques supporting the understanding and oversight of AI systems. However, explainability techniques are not necessarily the only means available to understand and interpret AI system outputs, and as such, it is not a requirement (and maybe not even technically feasible) that every high-risk AI system is explainable.

Nevertheless, less advanced techniques may also be highly effective and contribute to compliance with the requirements of the regulation, ranging from documentation approaches to the use of suitable user interfaces for the provision of information during system operation, as well as the presentation of well-calibrated

confidence measures. Suitably, IEEE P7001 defines levels of transparency with an increasing range of sophistication and complexity, considering in the first two levels documentation approaches, scenarios, principles of operations, as well as interactive training materials, well-aligned with Article 13 requirements for the provision of information to users. Some approaches prescribed in these initial levels, especially for domain expert users and super users, are also relevant towards fulfilling Article 14 requirements for human oversight, as they explicitly demand that the material provided allows for a rehearsal of interactions with the system and includes safe operation and supervision aspects.

The higher transparency levels defined in the IEEE work are already part of the domain of explainable AI, which, at the time of writing, is an active field of research. Explainable AI techniques, even though not a strict requirement for all high-risk AI systems, would enable high degrees of AI transparency and greatly support human oversight. Indeed, having a trustworthy explanation of the AI model decision-making process may be one of the most effective means to ensure that the human overseeing the system is "able to decide, in any particular situation, not to use the high-risk AI system or otherwise disregard, override or reverse the output of the high-risk AI system" as required in Article 14. Nevertheless, the field of explainable AI is still in its infancy and more research is needed to reliably integrate it into AI products. At the present moment there may be few state-of-the-art approaches in explainable AI suitable for standardisation, i.e. those techniques that are generally accepted as good practice. This is expected to make some of the higher levels of transparency defined in current standardisation work difficult to achieve in practice. This limitation applies certainly to the IEEE standard analysed, but also to similar ISO/IEC work, such as ISO/IEC TS 6254 “Objectives and approaches for explainability of ML models and AI systems”. This technical specification describes approaches and methods that can be used to achieve the explainability objectives of different stakeholders concerning the AI system's behaviour output and results, and identifies several characteristics of explainability (explanation needs, form, approaches, and technical constraints), using them to categorise existing approaches. However, an important limitation of the version of ISO/IEC TS 6254 reviewed is that it does not discuss nor compare the technological maturity and known limitations of the methodologies it covers. Indeed, as highlighted in the document: "for a number of explainability methods it is still an open research question whether the explanations they provide are representative of how the AI system arrived at the final decision". This means that some of the explanations provided by the explainable AI methodologies presented in ISO/IEC TS 6254 may not be trustworthy, i.e., they may not reflect the actual decision-making process of the AI system. Other known limitations of some explainable AI methodologies that would need to be discussed in ISO/IEC TS 6254 include: a lack of stability and robustness of explanations (i.e., the fact that different runs of one methodology might provide different explanations for the same instance), lack of explanation comprehensibility, and the negative impact of AI explanations on automation bias.

Despite this, and while many important aspects of explainable AI technologies are open research questions, they can be useful for expert stakeholders to understand their current limitations. This group of stakeholders, such as AI developers, can make use of explainable AI in the development phase of the model, e.g. as a debugging tool or to perform sanity checks throughout the model lifecycle. Indeed, explainable AI techniques can be helpful to test model robustness and accuracy and, therefore, can play a role in the context of Article 15 of the AI Act. This aspect is also reflected in the considerations on explainability of ISO/IEC TS 6254: “being able to explain a certain system behaviour during system testing helps AI developers to debug the system”.

Considering this, the standardisation of explainable AI methodologies targeted at expert users might be beneficial in the context of the Article 15 of the AI Act, and is currently expected to be covered by the ESOs in upcoming work, e.g. CEN-CENELEC-ETSI PWI “Overarching unified approach on trustworthiness characteristics”. As is the case with bias, the work on human oversight and explainable AI does not need to start from scratch. ESOs should consider adopting technical content from IEEE P7001 in the context of transparency and human oversight, in addition to the already discussed aspect of record keeping. This may require adaptations aiming to ensure that realistic transparency requirements in line with the AI Act are prioritised. European standardisers should also consider ISO/IEC TS 6254 as a future relevant reference. In fact, the scope of this document appears to be very complementary with IEEE P7001: while the IEEE standard sets out transparency desiderata without defining how to achieve them, the ISO/IEC specification describes explainable AI methodologies that might be able to satisfy some of them. In this sense, ISO/IEC TS 6254 provides a catalogue of explainability techniques which could be very informative for AI developers. However, its adoption in support of the AI Act should include a realistic assessment of their maturity, readiness and added value for concrete stakeholders (e.g. developers and/or users). And crucially, standardisation of explainability methods should not result in other relevant standardisation work in the context of human oversight being neglected. Some relevant areas which appear to be under-represented in the current AI standardisation landscape include technical specification covering mature, tested and robust methods to support users'

interpretation of AI outputs, even if simpler and not based on explanations, and other human oversight aspects such as controllability of AI systems.

5.3 Risk management and ethical design processes

AI risk management is an area covered by existing standards at the ISO/IEC level such as ISO/IEC DIS 23894, which provides guidelines for organizations to integrate risk management into their AI-related activities. It builds upon the risk management guidelines provided by ISO 31000, complementing them with AI-specific considerations. The generic risk management framework described in ISO/IEC 23894 could be adopted by organizations that develop products using AI, and remains, overall, in line with the requirements of Article 9(2) of the AI regulation, as it describes a risk identification, analysis and evaluation pipeline that can be systematically implemented throughout the product's lifecycle. The AI-specific considerations and guidelines that ISO/IEC 23894 provides on top of ISO 31000 include useful examples of what AI adopters and providers may need to take into consideration. Further relevant informative content can be found in the annexes, such as a preliminary list of AI risk sources, or a mapping of the risk management process to the AI lifecycle. However, ISO/IEC 23894's content is mostly informative, resulting in a document that provides useful guidelines and principles but is light in terms of formal specification that could be subject to compliance assessment. Furthermore, the focus of this document is to provide guidance on managing risks faced by organizations, and therefore, it does not appear to be specifically intended for risk management in the context of a concrete product development activity, e.g. involving a high-risk AI system, as defined in Article 9 of the AI regulation.

A complementary product-oriented view of risk management can be provided by the IEEE 7000 standard as it details a process to systematically consider and address ethical values and risks in the design of an AI system considering its context of use. The process covers the initial product design and development stages, and results in the definition of suitable ethical requirements that consider interactions of AI technology with individuals as well as their impact. These ethical requirements can be managed, assessed and validated alongside other requirements, e.g. technical ones, leading to concrete features and controls that can be implemented and tested in the AI system. This effectively results in a risk-based design process in line with the requirements of the AI Act, as well as with well-established system engineering practices. Therefore, it should be possible to integrate this standard with existing development processes used by a wide range of AI providers, representing a valuable input to developers of high-risk AI systems across different sectors and of different sizes in need of concrete, systematic development processes. Consequently, European Standardisation organizations could consider the integration of this standard, while guaranteeing its suitability to the needs of the AI Act. This would require, among other things, ensuring that the standard, which is currently neutral in respect to the values to be reflected in the design, prioritizes European values and the specific AI risks and protected interests emphasized in the AI Act, namely those related to the health, safety and fundamental rights of individuals.

For this and other reasons, such as its strong process orientation and partial coverage of the AI design and development lifecycle, the IEEE 7000 standard cannot cover all risk management requirements in the legal text in a stand-alone manner. Some complementary elements required are technical specifications covering concrete AI risk sources and providing checklists and guidelines for their assessment and mitigation. These are also not currently provided by ISO/IEC 23894, which in its turn is also lacking other key elements of Article 9, such as a detailed consideration of the interactions between different requirements for high-risk AI systems, communication and treatment of residual risks or the definition of the necessary scope, procedures and evaluation metrics for testing AI systems in the context of a risk management process. All of the aforementioned missing aspects are expected to be covered by new European standardisation work, such as CEN-CENELEC-ETSI PWI "Risk catalogue and risk management" recently proposed at the ESO level.

6 Discussion

While only including a relatively small sample of existing IEEE documents on AI standardisation, our analysis has identified a substantial amount of content that partially covers most of the requirements for high-risk AI systems under the AI Act. This is depicted in Figure 1, where the different shades capture the scores assigned by experts through qualitative analysis of each document against AI Act standardisation needs. These scores consider both breadth and depth of coverage of the respective articles in the legal text. It should be noted, however, that these values provide only an approximate measure of the relevance of these standards. For example, there are key technical considerations of AI systems, such as bias-related risks, that are not covered by a single article in the legal text, but are prevalent throughout the requirements for high-risk AI systems in the regulation. Consequently, technical specifications covering these aspects may have lower associated scores in the individual requirements but still fulfil critical standardisation needs. Furthermore, individual standards and certification criteria have been assessed only against the main legal requirements covered, even if in some cases their clauses included considerations partially touching on most or all requirements. The requirements not considered in the analysis due to their relatively minor coverage are represented as white shades in Figure 1. Given these considerations, the reader is advised not to solely rely on these scores, taking them as a guidance and referring to the full analysis of individual standards provided in sections 4 and 5.

The two types of documents reviewed, namely standards and certification criteria, are very different in nature. Standards tend to focus on a narrow set of requirements and concerns with a stronger depth. All of the standards reviewed are mature, applicable to a broad range of AI systems and well-aligned with the horizontal nature of the AI regulation. They are mostly process-oriented specifications, and are relevant to the AI Act. In particular, we identified 3 highly valuable standards: the IEEE 7000 –or forthcoming ISO/IEC/IEEE 24748-7000 after its planned adoption by ISO/IEC– Standard Model Process for Addressing Ethical Concerns during System Design; the IEEE P7001/D4 Draft Standard for Transparency of Autonomous Systems; and the IEEE P7003/D1 Draft Standard for Algorithmic Bias Considerations.

As discussed in section 5, technical specifications provided by some of these documents would support the operationalisation of AI Act requirements for high-risk AI systems, complementing the ISO/IEC landscape. In some cases, where adoption as a whole may not be the most effective way forward, our analysis could facilitate selection of relevant content to be integrated into future European specifications. Furthermore, we identify concrete aspects where this content should be complemented from the lens of the European AI Act. These include establishing stronger links to AI-specific risks and to state-of-the-art techniques, providing concrete metrics for measuring AI trustworthiness, or in some cases, ensuring coverage of the full AI system lifecycle. These recommendations are captured as part of our detailed analysis in section 4. In addition, when considering complementarities with other standards from ISO/IEC, an alignment of the terminology employed would also be needed, e.g. with regard to specific terms related to risk, transparency and bias.

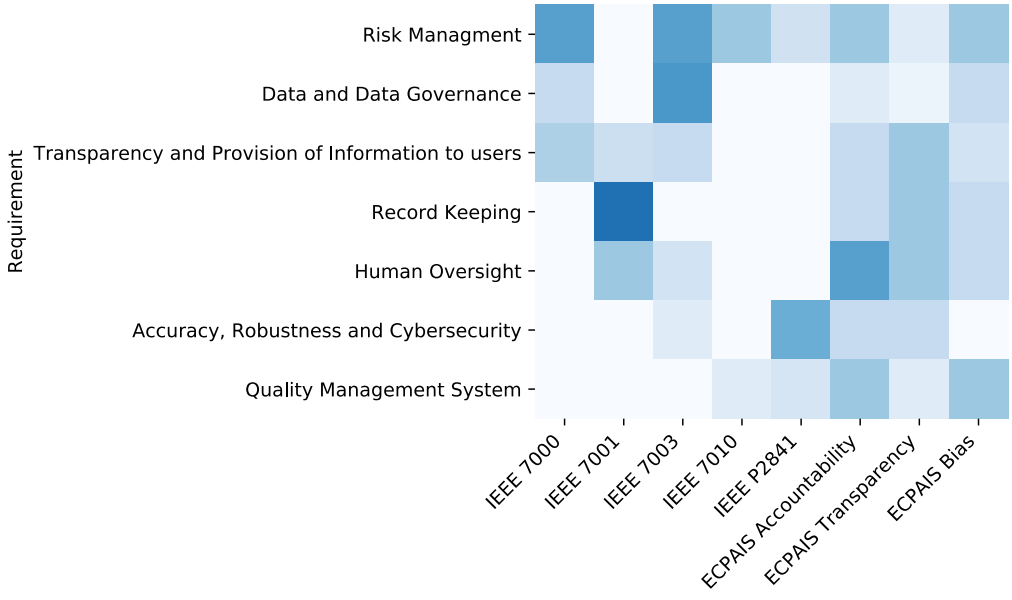


Figure 1 Overview of the coverage of legal requirements provided by the analysed standards and certification criteria

Certification criteria documents are broader than standards, and are targeted to conformity assessment stakeholders. As such, they contain large sets of outcome-oriented requirements and evidence covering elements from most articles in the legal text, albeit with less depth and without extensive detail on concrete methods and processes to achieve the desired outcomes. Given their extensiveness, they tend to score in most or all of the legal requirements, but at the same time include many criteria are outside the current scope of the AI Act.

Despite this, the evidence described in certification criteria suites is aligned with the quality management system and technical documentation described in the legal text as a basis for conformity assessment. As such, proper selection and consolidation of a subset of these criteria could form a basis for the assessment of high-risk AI systems. Naturally, this would require a prior careful alignment of the selected elements with the content of future standards supporting the AI regulation, something that may be challenging at this time, as many of them are still in early stages of development. Nevertheless, some recommendations are made throughout this document to make these certification suites more effective in the context of the AI Act, contributing to the need for implementable methods for verifying compliance with the future AI regulation. Concrete recommendations include providing further detail on how to assess and measure the evidence requested, specifying some it in a more concrete and objective manner. This aspect is expected to be further developed and adapted based on practical experience. Another important consideration would be the definition of a process to select and match criteria to the characteristics of specific AI systems as well as the size and resources of the AI system provider's organization.

Considering they are still in development, future iterations and additional documentation and certification programmes planned by the IEEE around these certification criteria could address many of these points. Indeed, since this review was carried out, the IEEE has made available an implementation concept for the certification criteria, the certifAIed certification program, including a harm-benefit analysis aimed at discovering the societal impact of AI products and producing a corresponding selection of the relevant suites of criteria. The certification process prescribes a standardised Case for Ethics documentation as a basis for an assessment that is not limited to binary pass/fail decisions, but enables assessors to make recommendations for improvement. In the context of the AI Act, this approach could in the future make the ECPAIS criteria relevant not just in the context of conformity assessment, but also as a process for providers to implement a continuous and iterative process to address ethical risks in their AI products. Given its potential to be applied and tailored to a wide range of use cases and domains, the complete certification material could be the subject of a future review.

Collectively, the standards and certification criteria reviewed provide significant coverage of requirements for high-risk AI systems defined in the AI Act, including some of those for which a relative scarcity of international standards has been observed, such as addressing bias-related risks, ensuring appropriate human oversight, or implementing record keeping mechanisms. In this regard, and considering their level of maturity, IEEE standards appear to fittingly complement the ISO/IEC standardisation landscape.

7 Conclusions

In this report, we present an in-depth analysis of several AI standards and certification criteria suites from the IEEE Standards Association. This analysis has been carried out by a group of experts in the field of Trustworthy AI from the European Commission's Joint Research Centre with the objective to assess the degree to which these specifications cover European standardisation needs in the context of the AI Act.

Overall, the documents reviewed have been found to provide relevant technical detail that could support providers of high-risk AI systems in complying with the requirements defined in the legal text. More importantly, some of the reviewed specifications focus on concrete technical areas which have been flagged as standardisation gaps by previous analyses, making them potentially valuable sources for the definition of European and harmonised standards for the AI regulation.

The provision of these standards, upon reception and acceptance of a request from the European Commission, is the remit of European Standardisation Organisations. In their capacity, they are able to leverage existing specifications, adapting them if required to the European regulatory context. Building on existing international work on AI is expected to be an efficient way to develop the standards needed for the AI Act, avoiding duplication of efforts and facilitating their broad adoption by AI providers. A primary source of relevant AI standards for adoption in the European context is ISO/IEC, a process facilitated by existing collaboration agreements. However, similar arrangements are also possible with other prominent SDOs such as the IEEE Standards Association. Indeed, there have been positive developments in this direction, including the launch of workshop agreements involving CEN-CENELEC, IEEE and other standardisers to jointly work on topics such as digital sovereignty. Another significant development is the re-establishment of a category-A liaison between IEEE and ISO/IEC JTC1 SC42, which may in turn open further opportunities for cooperation on AI standardization with CEN and CENELEC.

We see these as promising steps which strengthen the links between IEEE and European standardisers, paving the way towards possible future collaborations aiming to capitalize on the significant work undertaken by the IEEE Standards Association on Trustworthy AI, and to adjust it to the European context, adding further momentum to the substantial efforts underway to provide technical specifications in support of the AI Act.

References

[1] European Commission, "Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain union legislative acts.," 2021.

[2] European Commission, "New legislative framework," [Online]. Available: https://ec.europa.eu/growth/single-market/goods/new-legislative-framework_en. [Accessed 23 03 2022].

[3] European Commission, "Key players in European Standardisation," [Online]. Available: https://ec.europa.eu/growth/single-market/european-standards/key-players-european-standardisation_en. [Accessed 23 03 2022].

[4] International Organization for Standardization (ISO); European Committee for Standardization (CEN), *Agreement on technical co-operation between ISO and CEN (Vienna Agreement).*

[5] (IEC), International Electrotechnical Commission; (CENELEC), European Committee for Electrotechnical Standardization, *IEC - CENELEC Agreement on common planning of new work and parallel voting (Frankfurt Agreement).*

[6] S. Nativi and S. De Nigris, "AI Watch, AI standardisation landscape state of play and link to the EC proposal for an AI regulatory framework," Joint Research Centre (European Commission), 2021.

[7] StandICT.eu, "Report of TWG AI: Landscape of AI Standards," 2021.

[8] "Draft standardisation request to the European Standardisation Organisations in support of safe and trustworthy artificial intelligence," 2022. [Online]. Available: https://ec.europa.eu/docsroom/documents/52376.

[9] "ISO/IEC Joint Technical Committee 1 / Subcommittee 42 (Artificial intelligence)," [Online]. Available: https://www.iso.org/committee/6794475.html. [Accessed 23 03 2022].

[10] "IEEE Standards Association," [Online]. Available: https://standards.ieee.org/. [Accessed 23 03 2022].

[11] I. Hupont, M. Micheli, B. Delipetrev, E. Gómez and J. Soler Garrido, "Documenting high-risk AI: an European regulatory perspective," *TechRxiv,* 2022.

List of abbreviations and definitions

A/IS	Autonomous and Intelligent Systems
AI	Artificial Intelligence
CEN	European Committee for Standardisation
CENELEC	European Committee for Electrotechnical Standardization
EC	European Commission
ECPAIS	Ethics Certification Program for Autonomous and Intelligent Systems
ESO	European Standardization Organization
ETSI	European Telecommunications Standards Institute
EU	European Union
IEC	International Electrotechnical Commission
IEEE	Institute of Electrical and Electronics Engineers
ISO	International Organization for Standardization
ITU-T	International Telecommunication Union - Telecommunication Standardization Sector
JRC	Joint Research Centre
JTC	Joint Technical Committee
SC	Subcommittee
SDO	Standards Development Organization

List of figures

List of tables

GETTING IN TOUCH WITH THE EU

In person

All over the European Union there are hundreds of Europe Direct centres. You can find the address of the centre nearest you online (european-union.europa.eu/contact-eu/meet-us_en).

On the phone or in writing

Europe Direct is a service that answers your questions about the European Union. You can contact this service:

— by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),

— at the following standard number: +32 22999696,

— via the following form: european-union.europa.eu/contact-eu/write-us_en.

FINDING INFORMATION ABOUT THE EU

Online

Information about the European Union in all the official languages of the EU is available on the Europa website (european-union.europa.eu).

EU publications

You can view or order EU publications at op.europa.eu/en/publications. Multiple copies of free publications can be obtained by contacting Europe Direct or your local documentation centre (european-union.europa.eu/contact-eu/meet-us_en).

EU law and related documents

For access to legal information from the EU, including all EU law since 1951 in all the official language versions, go to EUR-Lex (eur-lex.europa.eu).

Open data from the EU

The portal data.europa.eu provides access to open datasets from the EU institutions, bodies and agencies. These can be downloaded and reused for free, for both commercial and non-commercial purposes. The portal also provides access to a wealth of datasets from European countries.

The European Commission's science and knowledge service

Joint Research Centre

JRC Mission

As the science and knowledge service of the European Commission, the Joint Research Centre's mission is to support EU policies with independent evidence throughout the whole policy cycle.

EU Science Hub
joint-research-centre.ec.europa.eu

@EU_ScienceHub

EU Science Hub - Joint Research Centre

EU Science, Research and Innovation

EU Science Hub

EU Science

Publications Office
of the European Union