Evaluating Dimensions of AI Transparency: A Comparative Study of Standards, Guidelines, and the EU AI Act

Sergio Genovesi
SKAD AG, Frankfurt am Main, Germany

Martin Haimerl
Universität Furtwangen, Germany

Iris Merget
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI), Kaiserslautern, Germany

Samantha Morgaine Prange
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI), Kaiserslautern, Germany

Otto Obert
Main DigitalEthiker GmbH, Karlstadt am Main, Germany

Susanna Wolf
DATEV eG, Nürnberg, Germany

Jens Ziehn
Fraunhofer IOSB, Karlsruhe, Germany

Abstract

Transparency is considered a key property with respect to the implementation of trustworthy artificial intelligence (AI). It is also addressed in various documents concerned with the standardization and regulation of AI systems. However, this body of literature lacks a standardized, widely-accepted definition of transparency, which would be crucial for the implementation of upcoming legislation for AI like the AI Act of the European Union (EU). The main objective of this paper is to systematically analyze similarities and differences in the definitions and requirements for AI transparency. For this purpose, we define main criteria reflecting important dimensions of transparency. According to these criteria, we analyzed a set of relevant documents in AI standardization and regulation, and compared the outcomes. Almost all documents included requirements for transparency, including explainability as an associated concept. However, the details of the requirements differed considerably, e.g., regarding pieces of information to be provided, target audiences, or use cases with respect to the development of AI systems. Additionally, the definitions and requirements often remain vague. In summary, we demonstrate that there is a substantial need for clarification and standardization regarding a consistent implementation of AI transparency. The method presented in our paper can serve as a basis for future steps in the standardization of transparency requirements, in particular with respect to upcoming regulations like the European AI Act.

2012 ACM Subject Classification Computing methodologies → Artificial intelligence

Keywords and phrases AI, transparency, regulation

Digital Object Identifier 10.4230/OASIcs.SAIA.2024.10

Category Academic Track

Acknowledgements We would like to thank our fellow members of the German Standardization Roadmap on Artificial Intelligence, supported by the Federal Ministry for Economic Affairs and Climate Action (BMWK) on the basis of a decision by the German Bundestag; the Foundations working group; the Ethics sub-working group; and especially the organizing committee of DIN, the German Institute for Standardization, and DKE, the German Commission for Electrical, Electronic & Information Technologies of DIN and VDE, for providing the basis for the activities of this working group and in particular this publication.

1 Introduction and Motivation

Transparency has been identified as one of the key features for trustworthy AI by the international expert community [1, 3, 4, 5, 6, 8, 10, 16, 21]. However, the existing body of literature – ranging from academic papers, policy documents, recommendations, to regulations, and standards – lacks a standardized, widely-accepted definition of transparency. The documents provide varying interpretations, focusing on different dimensions of transparency such as traceability of data origin, explainability of algorithmic decisions, disclosure of the system's abilities and limitations, and assuring user awareness about them interacting with a machine, among others. This fragmentation of interpretations leads to differing requirements for achieving transparency in AI systems, posing a significant challenge for standardization and compliance assurance with respect to AI quality.

This paper is a pilot study aiming at establishing a methodology to highlight both discrepancies and commonalities across key documents, providing a tool for policy makers to identify relevant features that need to be addressed to produce effective standards for AI transparency within a specific regulatory framework. For our analysis, we focused on selected standards and guidelines of high relevance within the European framework published by prominent German, European and international organizations during the drafting phase of the European AI Act. The AI Act itself was also included as an important reference.

By considering further documents representing other perspectives and fields of interests, our methodology has the potential to scale up to other theoretical, practical, and regulatory frameworks with differing geographical focus. This includes additional transparency dimensions being defined by other documents.

2 Considered Documents

We compared transparency definitions and requirements in several pivotal documents from the field of AI regulation and standardization shown in Tab. 1. Besides the AI Act itself [7], considered sources include central papers with respect to the development of the AI Act, i.e., the HLEG GL [6], and OECD [18], already available documents from standardization, i.e., ISO 22989 [12] and IEEE 7000 or [11], as well as further guidelines and internationally recognized white papers in this direction, i.e., VDE 90012 [22] and Fraunhofer GL [20]. The selection of the documents was based on discussions that were conducted in the context of the German Standardization Roadmap on Artificial Intelligence [4] and its subsequent activities. According to its character as a pilot study, this paper did not include a comprehensive analysis of potentially relevant documents in this field, but focused on this specific selection. In the following, we present the documents in detail.

2.1 HLEG GL: HLEG Ethics Guidelines for Trustworthy AI

The "Ethics Guidelines for Trustworthy AI" by the Independent High-Level Expert Group on Artificial Intelligence of the European Commission [6] is a report published in 2019 defining a European framework for achieving trustworthy AI. These guidelines revolve around ensuring AI systems are lawful, ethical, and robust, commencing from their development to their deployment and operation. They put forward a set of seven key requirements including human agency and oversight, technical robustness and safety, privacy and data governance, transparency, diversity and fairness, societal and environmental well-being, as well as accountability. A notable part of the guidelines is a detailed assessment list designed to guide practical implementation.

Table 1 Overview of the documents considered in this study, by used abbreviation, official title as appears on the document, and corresponding handle to the entry in the references section.

Abbv.	Official document title	Reference
HLEG GL	Independent High-Level Expert Group on Artificial Intelligence (HLEG) set up by the European Commission: Ethics Guidelines for Trustworthy AI	[6]
AI Act	European Parliament legislative resolution of 13 March 2024 on the 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 (Texts Adopted)	[7]
ISO 22989	ISO/IEC 22989:2022: Information technology — Artificial intelligence — Artificial intelligence concepts and terminology	[12]
OECD	OECD Framework for the Classification of AI Systems	[18]
VDE 90012	VCIO based description of systems for AI trustworthiness characterisation VDE SPEC 90012 V1.0 (en)	[22]
Fraunhofer GL	Fraunhofer IAIS: Guideline for Designing Trustworthy Artificial Intelligence – AI Assessment Catalog	[20]
IEEE 7000	IEEE Std 7000-2021: IEEE Standard Model Process for Addressing Ethical Concerns during System Design	[11]

2.2 AI Act: European Artificial Intelligence Act

The "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" [7], commonly referred to as the "European AI Act", is a comprehensive legal framework on AI devised by the European Union. Representing the first such framework worldwide, it aims to foster trustworthy AI within Europe and beyond by ensuring that AI systems uphold fundamental rights, safety, and ethical standards. Addressing the diverse impact of AI systems, the Act categorizes AI technologies into different risk levels. AI systems carrying an unacceptable level of risk are prohibited. High-risk AI systems face stringent requirements to manage their risks, including issues related to transparency. For limited-risk applications, the AI Act prescribes specific transparency obligations, ensuring an informed and aware interaction with the AI system. Furthermore, the Act outlines particular transparency obligations for all general-purpose AI (GPAI) models and more specific requirements for GPAI models with systemic risk. Additionally, the AI Act imposes stringent obligations on all actors in the AI value chain, ranging from providers and deployers to importers and distributors, among others, ensuring a rigorous approach to enforcement and compliance across the European market.

2.3 ISO 22989

ISO 22989 [12] "Information technology — Artificial intelligence — Artificial intelligence concepts and terminology", released in 2022, aims to establish a common terminology and concepts for the field of AI with a very general audience.

Terms ranging from "AI agent" to "validation data" are structured into seven categories and defined briefly, usually in single-line descriptions, while the descriptions of concepts span one to several paragraphs each, split into 19 categories. Additionally, elements such as the AI life cycle or AI ecosystems are defined and explained.

2.4 OECD: Framework for the Classification of AI Systems

Based on the first version of the OECD AI Principles [17], the OECD Framework for the Classification of AI Systems [18] is a tool designed to help policymakers, regulators, and legislators characterize AI systems for aligned policy action. This framework examines the spread of AI across sectors, recognizing the variations in benefits, risks, and policy challenges offered by different AI system types. By highlighting system characteristics critical for technical and procedural measure implementation, it aims at facilitating policy debate, supporting risk assessment, and helping in developing AI-related policies and regulations. The framework is structured along five key dimensions, including People & Planet, Economic Context, Data & Input, AI Model, and Task & Output, each with sub-dimensions important for policy considerations. It also distinguishes between AI "in the lab" and AI "in the field," offering a baseline for promoting common AI understanding, informing AI registries, and supporting sector-specific frameworks, risk assessment, and management throughout the AI system life cycle.

2.5 VDE 90012: VCIO Based Description of Systems for AI Trustworthiness Characterisation, VDE SPEC 90012

The VDE SPEC 90012 "VCIO based description of systems for AI trustworthiness characterisation" by the German Association for Electrical, Electronic & Information Technologies (VDE) [22] provides a framework for describing socio-technical attributes of systems with integrated AI, particularly where high levels of trust are required. It explains the VCIO (Values Criteria Indicators Observables) model, which evaluates a product's adherence to specific values and its trustworthiness, potentially supporting a trust label certification. This characterization is versatile, serving end consumers, companies, and government entities for setting requirements or comparing products. The assessment allows for different values such as privacy and transparency, and supports tailoring target requirements during product development for value compliance. Notably, while independent of the product's risk level and without setting minimum standards, the description aligns with the European AI Act, offering a delineation of trustworthiness that demonstrates compliance and market differentiation. Focusing on AI-specific features like datasets, scope, processes, and responsibilities, the standard also encompasses broader elements essential for establishing AI trustworthiness. This VDE SPEC aims to enable a reproducible and transparent classification of AI systems according to their degree of fulfillment of values or competencies, and to allow an assessment of the extent to which the requirements for achieving a certain risk level are met.

2.6 Fraunhofer GL: Fraunhofer IAIS Guideline for Designing Trustworthy Artificial Intelligence

The "Guideline for Designing Trustworthy Artificial Intelligence" released by Fraunhofer IAIS [20] provides a structured approach to define application-specific assessment criteria emphasizing quality and trust as competitive advantages. This guideline is targeted at data scientists in the development stage, and assessors in quality assurance for AI applications. It outlines a four-step assessment process encompassing comprehensive risk analysis, setting measurable targets, listing measures to achieve those targets, and establishing a safeguarding argumentation. The guideline focuses on six dimensions of trustworthiness: fairness, autonomy and control, transparency, reliability, safety and security, and data protection. It includes established KPIs to quantify targets and offers guidance on the documentation of technical and organizational measures reflective of the current state of the art to mitigate AI-related risks.

2.7 IEEE 7000: IEEE Standard Model Process for Addressing Ethical Concerns during System Design

The IEEE Std 7000™-2021: IEEE Standard Model Process for Addressing Ethical Concerns during System Design [11] was released in 2021 and aims to standardize approaches to consider ethical aspects in the development of systems. As indicated by its title, IEEE 7000-2021 focuses on the development process of the system rather than on properties of the system itself. It is also not exclusively an AI-related standard, describing itself as "applicable to all kinds of products and services, including artificial intelligence (AI) systems". The five year development process of the standard means that it is not to be understood as a response to very recent advances in generative AI and large language models. The primary audience is "engineers and technologists" viewed from the organization level for whom a set of processes is proposed (including an Ethical Values Elicitation and Prioritization Process, an Ethical Risk-Based Design Process and a Transparency Management Process) to enable them to include ethical aspects into the system development.

In the collection of documents cited here, IEEE 7000-2021 stands out as the only document not primarily conceived for AI applications but for systems development in general, which also reflects onto its perspective of transparency.

3 Comparison Criteria: "Does the document distinctly..."

In this section, we define and delimit our evaluation criteria. The goal of the process was to achieve a comparison that indicates whether or not the documents incorporate or express a specific notion or scope of transparency. Based on the criteria defined below, the particular documents were evaluated. The evaluation had three possible results:

1. yes – the document distinctly adopts the notion of transparency expressed by the criterion;
2. no – the document does not refer to the notion of transparency expressed by the criterion;
3. unclear – the document does not explicitly introduce the notion of transparency expressed by the criterion, but may contain implicit support or alignment with it.

The criteria were developed according to discussions conducted as a follow-up of the German Standardization Roadmap on Artificial Intelligence [4]. This was performed in an iterative approach where relevant aspects were collected from the included documents and the criteria were consolidated accordingly before the evaluation process was started. The authors consider the list of criteria deduced by the analysis of the above-mentioned documents to be exhaustive, meaning that no other additional generic transparency criteria were found in the reviewed documents. The authors do not exclude that new transparency criteria can be deduced, by analyzing further documents. This is part of the iterative approach and the authors recommend screening for additional transparency criteria when expanding this evaluation methodology to an extended set of documents.

For each paper, at least two members of the authors group performed the analysis. In case of a disagreement, the particular rating was discussed in the overall group of authors, who finally decided about the classification. For practical reasons, not all authors could be involved every time. For achieving a reliable consensus, at least five of the authors had to be involved into the final decision where the two authors which analyzed the document needed to be included. In cases where no full agreement could be achieved, the item was assigned to the "unclear" category. Note, however, that the evaluation category "unclear" refers to the way a specific transparency criterion is presented within a document and not to the modalities of agreement among authors.

The following list describes the used criteria in a systematic way. All criteria start with the phrase "Does the document distinctly...", followed by the criterion, such as "... set transp. requirements relating to the design or development stage of AI?".

The term "distinctly" as opposed to "explicitly" is chosen deliberately to include clearly implied intentions, since the choice of words and voice differs considerably between documents. For example, IEEE 7000 [11] states:

> System requirements for machine learning systems may include quantitative and qualitative data-oriented specifications that include identifications for collection of data, data formats, diversity, ranges of data, ...

This clearly implies that IEEE 7000 considers transparency requirements relating to the design and development stage of AI, since most commonly for machine learning systems, data will be used for training and testing during these stages – even if no explicit reference to the design and development stage of the life cycle is made. Such clearly implied intentions are, for this study, considered equivalent to explicit statements, with particular decision principles given in the following subsections.

In this process, it should be understood that the task of comparing documents that widely differ in focus, authors' background, scope and intended impact cannot be strictly formal and performed under sharply defined criteria while still extracting meaningful results. The goal of this approach is to adequately reflect the conceptual ideas incorporated into the respective documents, requiring to some extent a margin of discretion and interpretation. The same should be applied by the readers who are to realize that the following will not replace a formal, thorough study of any individual document, possibly with technical and legal expertise, in particular when, for example, assessing a system with respect to a given standard or legal regulation, such as the AI Act. Furthermore, this motivates the aforementioned choice of three instead of just two possible evaluation results, namely "yes" and "no" when no considerable ambiguity was found, and "unclear" else (cf. Tab. 3).

For the description of the requirements, we use the coding scheme laid out in Tab. 2. For example, the code `TR-STG-OPS` ("Does the document distinctly set transp. requirements relating to the operation stage of AI?") is composed of the supergroup `TR` for "transparency", the group `STG` for life cycle stage-related criteria, and the criterion `OPS` for the operation stage.

3.1 ... define the term "transparency" or a closely related concept?

This criterion is satisfied if the document clearly attempts to define or delimit what the term "transparency" means (at least for the specific purpose of the document). Since terms such as "explainability" or "traceability" are frequently used interchangeably, a definition of one of these terms also satisfies the criterion if the document uses the alternative term in a closely related sense.

3.2 ... set transp. requirements relating to the design or development stage of AI?$^{\texttt{TR-STG-DDV}}$

This criterion is satisfied if the document advocates transparency requirements that must be met or at least considered during the design or development stages of the AI system life cycle, such as the disclosure of training data or AI model details.

Table 2 Abbreviations of the form `TR-XXX-YYY` used here to classify grouped transparency criteria.

Abbrv.	Does the document distinctly...
TR-	*Group: Transparency-related (always given here)*
STG-	*Group: Related to stages within the AI life cycle*
DDV	...set transp. requirements relating to the design or development stage of AI?
OPS	...set transp. requirements relating to the operation stage of AI?
EOL	...set transp. requirements relating to post-end of life/retirement/disposal stage of AI?
SYS-	*Group: Related to the AI system*
MDL	...relate transp. to technical AI properties, such as code or ML models?
WGT	...relate transp. to ML "weights" or "features"?
OUT	...relate transp. to explainability of particular outputs?
CSQ	...relate transp. to predictability of consequences?
LIM	...relate transp. to limits or error/failure modes of AI?
SOA	...limit requirements to a current "state of the art"?
DAT	...relate transparency to training data?
PII	...relate transp. to user data and/or privacy?
PRP	...relate transp. to an intended purpose of the AI?
BIZ	...relate transp. to business models/operator interests?
RVL	...require for transp. revealing to users that the system uses AI as such?
AUD-	*Group: Related to the target audience of the transp.*
SPC	...consider transparency to be target audience-specific?
DEF	...define one or more such target audiences?
USR	...name users as a target audience?
NNU	...name affected non-users as a target audience?
OPR	...name operators as a target audience?
TST	...name testing and auditing organizations as a target audience?
REG	...name regulators or authorities as a target audience?
DEV	...name developers and (direct) partners in the dev. process as a target audience?
DSA	...name other manufacturers/providers of downstream applications as a target audience?

3.3 ... set transp. requirements relating to the operation stage of AI?TR-STG-OPS

This criterion is satisfied if the document advocates transparency requirements that must be met or at least considered during the operation stage of the AI system life cycle, such as informing users that the system they are using is based on AI.

3.4 ... set transp. requirements relating to post-end of life/retirement/disposal stage of AI?TR-STG-EOL

This criterion is satisfied if the document describes transparency risks or requirements that address the end of life or post-end of life stage of AI, for example measures to be taken during the decommissioning of an AI system, such as assuring and documenting that all data and learned features have been deleted. This criterion explicitly does not refer to transparency problems during operation causing a decommissioning (this would be requirements for the operation stage), but only to transparency-related measures that must be taken once the decommissioning is decided.

It should be noted that no agreement exists whether the (post) end of life stage is part of the "AI life cycle" in general. Prominently, the OECD [19] explicitly does not include such a stage in its life cycle, stating:

> AI system lifecycle: AI system lifecycle phases involve: i) 'design, data and models'; which is a context-dependent sequence encompassing planning and design, data collection and processing, as well as model building; ii) 'verification and validation'; iii) 'deployment'; and iv) 'operation and monitoring'.

A similar subdivision is given in [17] and adopted in [18]. In contrast, ISO 22989 [12] and IEEE 7000 [11] do include this stage, the latter taking its life cycle definition from ISO/IEC/IEEE 12207 [15], as:

> **2.31**
> **life cycle**
> evolution of a system, product, service, project, or other human-made entity from conception through retirement

3.5 ... relate transp. to technical AI properties, such as code or ML models?TR-SYS-MDL

This criterion is satisfied if the document states that transparency may include insight into technical system properties such as code, algorithms or ML models – for example by requiring that corresponding design choices must be documented or that the code must be disclosed to certain parties. This criterion does not relate to the machine learning variables in the model – this is covered in **TR-SYS-WGT**. For some parameters (namely "hyperparameters" in machine learning), the distinction is not strict – however, this ambiguity did not arise in the review of the particular documents presented here.

3.6 ... relate transp. to ML "weights" or "features"?TR-SYS-WGT

This criterion is satisfied if the document advocates to disclose, for machine learning-based systems, parameters that were established through model training. These are typically referred to as weights or features.

3.7 ... relate transp. to explainability of particular outputs?TR-SYS-OUT

This criterion is satisfied if the document relates transparency to the provision of explanations for one particular output of the system. For example, in a system that identifies cancer cells in medical images, this could mean to additionally visualize relevant image regions and/or provision of similar reference images to either the professional operator, or the patient.

3.8 ... relate transp. to predictability of consequences?TR-SYS-CSQ

This criterion is satisfied if the document relates transparency to the possibility of predicting and limiting consequences of AI system outputs in its real-world application. This includes transparency with respect to residual risks and potential harms the AI system has. For example, a document may require a company selling an autonomous shuttle to indicate accident risks that can arise from an AI error. We distinguish between this criterion addressing the practical consequences (including long-term and indirect effects) and the criterion **TR-SYS-LIM**, which relates only to a description of immediate technical failure modes and limitations, e.g., in performance, without an actual estimation of the impact associated with the failure.

3.9 ... relate transp. to limits or error/failure modes of AI?TR-SYS-LIM

This criterion is satisfied if the document relates transparency to the disclosure of known limitations and error/failure modes of an AI system. This can include a description of known conditions under which the AI system cannot achieve the required performance (e.g., by specifying the operational design domain, ODD, cf. [2]) or a description of error rates. We distinguish between this criterion addressing the immediate technical failure modes on the technical level, and the criterion **TR-SYS-CSQ**, which relates to a description of practical and possibly physical and/or long-term consequences (primarily on the level of the real-world application).

3.10 ... relate transparency to training data?TR-SYS-DAT

This criterion is satisfied if the document relates transparency of a machine learning system to the disclosure of training, validation or testing data used in the design or development stages of the system – regardless of whether the document advocates disclosing training datasets completely, or documenting selected properties such as fairness or scale of the dataset(s).

3.11 ... relate transp. to user data and/or privacy?TR-SYS-PII

This criterion is satisfied if the document relates transparency to the disclosure of how the AI system utilizes, stores and/or shares user data, privacy-critical information, or personally identifiable information (PII).

3.12 ... limit requirements to a current "state of the art"?TR-SYS-SOA

This criterion is satisfied if the document limits the imposed transparency requirements to what is feasible or accepted based on the current state of the art or the best available techniques (BAT). This implies that the state of the art may evolve and lead to different requirements, but also that acceptance criteria must not be applied in hindsight of later developments. Additionally it implies that the current state of the art is likely to provide an acceptable result at the given time.

3.13 ... relate transp. to an intended purpose of the AI?TR-SYS-PRP

This criterion is satisfied if the document relates transparency to the disclosure of an "intended purpose" of the AI system, if such a purpose exists. The criterion serves to fix the application context as a basis for further development steps like risk management, but also transparency requirements concerning, e.g., which use context needs to be considered when providing information to users. The criterion **TR-SYS-PRP** may serve to convey system capabilities to users and avoid misunderstandings about its proper use. Beyond this, the disclosure may serve to reveal hidden interests – however, the particular case of hidden *business interests* is addressed by **TR-SYS-BIZ** specifically.

3.14 ... relate transp. to business models/operator interests?TR-SYS-BIZ

This criterion is satisfied if the document relates transparency to the disclosure of business interests of the suppliers or operators of the AI system, or similar interests for providing or operating the system. For example, the document may advocate that a company offering an app for health advice must disclose to users if their business interest is to build a general human health prediction model for insurance companies – even if the individual users' privacy

is believably protected in the process. Note that this criterion goes beyond **TR-SYS-PRP** in the sense that a document satisfying **TR-SYS-BIZ** must distinctly consider business interests (e.g., the collection of health data) to extend beyond the immediate purpose (e.g., the provision of a health advice app). To satisfy this criterion thus also means to acknowledge a possible conflict between transparency and business interests.

3.15 ... require for transp. revealing to users that the system uses AI as such?[TR-SYS-RVL]

This criterion is satisfied when the document requires or proposes for the benefit of transparency that users should be informed about the fact that the system uses or is based on AI, or that its output is (at least partially) AI generated. This criterion is only satisfied if the document clearly considers the fact itself important to reveal proactively. It is not, for example, satisfied if the document requires an AI system to adhere to regulations by which an interested user may come to know that it uses AI, for example by registering in an official database.

3.16 ... consider transparency to be target audience-specific?[TR-AUD-SPC]

This criterion is satisfied if the document indicates that transparency cannot always be defined universally, but instead should be evaluated with respect to, or designed for, a particular target audience. For example, this audience may be defined through professional experience, (lack of) AI literacy, personal handicaps, the specific role in the AI value chain respectively the usage of the AI system, or the level information relevant to the particular audience. This criterion is only satisfied if the document indicates that different levels of information or different means of presentation are adequate for different audiences. Merely listing different receiving groups or entities will not satisfy this criterion. Furthermore, the document must clearly suggest providing multiple such levels of information regarding the same system. A document that acknowledges the existence of different levels of expertise, but derives from this the requirement to release a single documentation that must be understandable to all stakeholders alike (e.g., by assuring all explanations can be understood by laypersons) will not satisfy this criterion.

3.17 ... define one or more such target audiences?[TR-AUD-DEF]

This criterion is satisfied if the document considers transparency to be target audience-specific (**TR-AUD-SPC**) and, in addition to this, names or defines concrete groups that may require different variants of transparency. It must not define detailed requirements for transparency, but it should specifically mention one or more target audiences with particular requirements.

3.18 ... name users as a target audience?[TR-AUD-USR]

This criterion is satisfied if the document defines one or more specific target audiences for transparency (**TR-AUD-DEF**) and lists, among these, the end users of the AI system.

3.19 ... name affected non-users as a target audience?[TR-AUD-NNU]

This criterion is satisfied if the document defines one or more specific target audiences for transparency (**TR-AUD-DEF**) and lists, among these, people affected by the AI system who are not actively involved in their operation (such as operators or users) in the sense of the economic concept of "negative externalities" [9, Chapter 1, p. 5; Chapter 5, p. 125 ff.].

Prototypical examples include pedestrians, who are affected by the operation of an AI-based automated road vehicle; or job candidates whose application documents are rated through an AI-based system used by recruiters.

3.20 ... name operators as a target audience?TR-AUD-OPR

This criterion is satisfied if the document defines one or more specific target audiences for transparency (**TR-AUD-DEF**) and lists, among these, the operators of the AI system when they differ from the users. We use the term "operator" in the sense of the "natural or legal person, public authority, agency or other body using an AI system under its authority except where the AI system is used in the course of a personal non-professional activity".[1] In particular, the "operator" in this case is the body directly responsible for the continued operation of an AI system.

3.21 ... name testing and auditing organizations as a target audience?TR-AUD-TST

This criterion is satisfied if the document defines one or more specific target audiences for transparency (**TR-AUD-DEF**) and lists, among these, officially appointed public or private testing and auditing organizations, registration centers, etc., who are responsible for performing particular tests of the AI system, certifying it, providing an operating license, or similar. This includes notified bodies or other conformity assessment bodies included into the certification or conformity assessment of an AI system.

3.22 ... name regulators or authorities as a target audience?TR-AUD-REG

This criterion is satisfied if the document defines one or more specific target audiences for transparency (**TR-AUD-DEF**) and lists, among these, regulators, authorities, legislators, or other public bodies responsible for controlling the development and operation of AI systems.

3.23 ... name developers and (direct) partners in the dev. process as a target audience?TR-AUD-DEV

This criterion is satisfied if the document defines one or more specific target audiences for transparency (**TR-AUD-DEF**) and lists, among these, persons or organizations who are involved in the development or production of the AI system. This includes developers and other departments in the own company, but also partners or suppliers along a supply chain who are involved in the development process and related areas.

3.24 ... name other manufacturers/providers of downstream applications as a target audience?TR-AUD-DSA

This criterion is satisfied if the document defines one or more specific target audiences for transparency (**TR-AUD-DEF**) and lists, among these, other manufacturers or providers who intend to use the system in a downstream application. That is, these entities modify, extend, or adapt the original system and thus, create a new system, e.g., with a modified scope

[1] This definition is verbatim from the AI Act [7], which, however, assigns the term "deployer" to it; whereas the AI Act regards "operator" as "a provider, product manufacturer, deployer, authorised representative, importer or distributor."

Table 3 Results of the comparison. ✓ indicates that the document distinctly adopts the concept of the criterion; – indicates that it does not. Cases where the distinction is unclear are marked with ○.

	HLEG GL	AI Act	ISO 22989	OECD	VDE 90012	Fraunhofer GL	IEEE 7000
Definition	✓	✓	✓	○	○	○	✓
TR-STG-DDV	✓	✓	✓	✓	✓	✓	✓
TR-STG-OPS	✓	✓	✓	✓	✓	✓	✓
TR-STG-EOL	–	–	✓	○	○	–	–
TR-SYS-MDL	✓	✓	✓	✓	✓	✓	○
TR-SYS-WGT	–	–	✓	✓	○	✓	–
TR-SYS-OUT	✓	✓	✓	✓	✓	✓	–
TR-SYS-CSQ	✓	✓	–	✓	○	✓	✓
TR-SYS-LIM	✓	✓	✓	○	○	✓	○
TR-SYS-SOA	✓	✓	–	–	○	✓	✓
TR-SYS-DAT	✓	✓	✓	✓	✓	✓	✓
TR-SYS-PII	○	✓	✓	✓	✓	✓	✓
TR-SYS-PRP	✓	✓	✓	✓	–	✓	✓
TR-SYS-BIZ	✓	–	○	○	–	✓	✓
TR-SYS-RVL	✓	✓	–	✓	✓	✓	○
TR-AUD-SPC	✓	✓	✓	✓	✓	✓	✓
TR-AUD-DEF	✓	✓	–	✓	○	✓	–
TR-AUD-USR	✓	✓	✓	✓	✓	✓	✓
TR-AUD-NNU	✓	✓	–	✓	✓	✓	✓
TR-AUD-OPR	○	✓	○	✓	✓	✓	✓
TR-AUD-TST	✓	✓	○	–	✓	✓	○
TR-AUD-REG	–	✓	○	–	✓	–	–
TR-AUD-DEV	○	✓	✓	–	✓	✓	✓
TR-AUD-DSA	○	✓	–	–	–	○	–

or intended purpose. For example, this applies when the manufacturer of a downstream application builds its system on a general purpose AI (GPAI) model by adapting this base model towards a specific use case. The question here is whether the provider of the GPAI model is requested to provide certain transparency information to the manufacturer of the downstream application. This criterion, based on the concept of GPAI, considers "downstream applications" to be potentially unforeseen by the developers of the original AI system; in contrast, `TR-AUD-DEV` refers to informing parties working along the same known supply chain.

4 Results

Table 3 provides an overview of our analysis of criteria for AI transparency dimensions across the documents described above. The key findings from our review are presented, subsequently. For each entry, a final consensus could be found in the group of authors, according to the defined evaluation process. In some cases, the discussions showed that an ambiguity in the

interpretation of the requirements persisted. Thus, the particular criterion had to be rated as "unclear", since the definitions in the corresponding document were not clear enough or could not be interpreted in a sufficiently consistent way.

Transparency is a subject of discussion in all of the documents, emphasizing its importance in AI development. However, in some documents its definition is unclear or not provided, i.e., in OECD, VDE 90012, and the Fraunhofer GL.

Explainability of system outputs is included as an essential dimension of transparency in nearly all the documents. The only document not mentioning this aspect, namely IEEE 7000, is also the sole source in our collection not specific to AI but to system design in general. In certain documents, such as the Fraunhofer GL, HLEG GL, and the AI Act, traceability and communication are additionally mentioned as integral parts of transparency.

Transparency is explicitly demanded during both the development and operational stages across nearly all documents, except for OECD. The end of life stage, in particular in terms of transparency obligations as a task in the retirement stage, are only made explicit in ISO 22989. Furthermore, no concrete measures are expressed for this stage, even when this stage was included in the document. As previously stated, it should also be noted that the inclusion of this stage as part of the life cycle is disputed.

All documents claim that transparency requirements must be target group-specific. However, the definition of target groups is often unclear or limited.

There are significant differences regarding which target groups and use cases are considered. End users are consistently included, and also indirectly affected stakeholders / non-users are explicitly mentioned in nearly all documents, with the exception of ISO 22989. Transparency towards operators is widely addressed as well, although not always explicitly, i.e., in the HLEG GL and ISO 22989. Transparency towards developers is absent in OECD and addressed unclearly in the HLEG GL. For the other documents, adequate information needs to be provided to the developers in order to achieve sufficient transparency. Transparency for testing or auditing organizations is explicitly mentioned just in four out of seven documents, while transparency for regulators and public authorities is only addressed in a dedicated way in the AI Act and VDE 90012. More complex scenarios for the implementation of AI systems involving different actors are usually not considered. This means that most of the documents refer to a situation with a single manufacturer of the AI system. Only the AI Act explicitely discusses more complex scenarios, where, e.g., a general purpose AI (GPAI) model is developed by one manufacturer and then integrated into a downstream application by another manufacturer. In these cases, the AI Act includes transparency obligations to be fulfilled by the manufacturer of the GPAI model.

Notable differences were recognized regarding the specific kind of information that should be made transparent. Transparency concerning code, models, and training data is consistently addressed across documents. As already mentioned, transparency with respect to the explanation of outcomes is required in most of the documents, except IEEE 7000. Similarly, information regarding an AI system's intended purpose and its predictable consequences is considered a significant transparency aspect in nearly all the documents, with the exception of VDE 90012 regarding the intended purpose and of ISO 22989 regarding the predictability of consequences. Information concerning the limitations and error modes of AI systems is generally addressed in the examined sources, even though some documents do not explicitly present this as a dimension of transparency. Finally, other aspects such as the reference to the state of the art, to the business model and operator interests, as well as the disclosure of technical parameters of the system (e.g., "weights" and "features") are not mentioned as transparency obligations in many of the documents.

Most of the documents also require that an AI system should reveal that the user is interacting with an AI system. This requirement is only absent in ISO 22989 and unclear in IEEE 7000. Finally, it can be recognized that many entries remain marked with a grey circle. This shows that a number of topics remains ambiguous in the documents.

5 Discussion

The presented results underline that transparency is regarded as an important ethical value in AI regulation and standardization literature. Also, there is a wide range of criteria that are considered as important requirements for AI transparency across different regulatory frameworks and guidelines. This variety of criteria reflects the fact that AI transparency is thought to have many dimensions, each of which demanding specific, tailored requirements to be addressed. Even though many transparency dimensions address overlapping technical or societal aspects, they allow for independent implementation of transparency measures. For example, informing users that they are engaging with an AI system, explaining to auditing organizations or authorities how the system processes inputs to reach outputs, and providing access to information about the AI model and training data can be executed separately using distinct approaches.

When considering the full extent of what defines transparency within the respective documents, considerable disparities become evident. No two documents share a completely compatible conception of this ethical value. A contributing factor is that the considered documents define the criteria for transparency – to a large degree – implicitly through the measures proposed, rather than comprehensively specifying concepts and goals first and deriving adequate measures systematically. Moreover, different kinds of documents have different scopes and focuses; therefore they tend to emphasize transparency aspects that are more relevant for their purposes. For instance, some of them include more technical specifications regarding appropriate measures of system transparency, while others, like the EU AI Act, focus more on high-level requirements related to AI transparency, explicitly leaving the task of defining concrete technical measures to domain-specific, technical standards.

The analyzed documents reflect that transparency basically deals with the question concerning which pieces of information should be provided to which stakeholder to deploy AI safely and responsibly. Indeed, transparency is generally considered a driver for trustworthiness, enabling meaningful and informed human interaction with the system. In order to do so, it is necessary to consider the system's intended purpose, which defines the specific technical solutions, user groups, and use cases for the AI system. The provision of information needs to be aligned with the expertise of the involved persons as well as their particular needs. For example, there is a considerable difference between laypersons, technical experts, and actors with specific domain, application or regulatory knowledge. Additionally, the adequacy of information depends on the particular role of the actors. For example, a developer needs different information in comparison to a user, provider, auditing organization, or public authority.

This also goes for cases where a manufacturer includes an AI model from a third party provider in a new downstream application. In particular, such more complex scenarios were intensively discussed in the legislative phase of the AI Act [7] in the context of AI models without a specific / narrow intended purpose. The AI Act defines these types of models as general purpose AI (GPAI) models and describes obligations, which information has to be made available by the third-party provider of the GPAI model in order to achieve sufficient transparency for the manufacturer of the downstream application. This substantially extends

the requirements, since the ethical impact can usually only be rated when the context of the application is clear. Thus, a new line of discussions was addressed in the AI Act in this regard compared to the other documents. This was due to the fact that most of these documents were older, i.e., written between 2019 and 2022, where high-impact large language models and other generative AI systems were not yet on the market.

The latest developments of the debate around transparency for GPAI models exemplify the highly dynamic nature of AI technology, which is leading to challenges regarding AI standardization and regulation. As an ethical value guiding the deployment of AI systems, transparency is a fundamental prerequisite for the achievement of other ethical principles. Indeed, a lack of awareness of the system impact and a limited understanding of its capabilities and limitations could hinder different actors along the AI value chain in the responsible use of a system, causing ethical issues ranging from safety risk to data misuse or group discrimination. Moving from these premises, the development of the AI Act [7] and the associated New Legislative Framework of the European Union are indicative of the goal to establish clearly-defined, internationally accepted rules and standards to enable ethical, value-centered use and implementation of AI systems. To this end, the exact definition of major terms, the operationalizability of concepts, and an adequate translation into concrete requirements for all relevant stakeholders along the AI value chain plays a pivotal role. The dimensions and criteria presented in the current paper are considered to serve as a starting point for systematically collecting and comparing requirements for AI transparency. In particular, this may apply to the identification of discrepancies in current standardization documents and guidelines as well as to a systematic compilation of key aspects in future development steps.

At the same time, the presented study has limitations to be considered. For practical reasons, the study was performed as a pilot study and limited to seven well-known documents in the AI context. However, a broader look that also includes scientific positions could help to establish a more thorough perspective in this very dynamic field. Additionally, the currently developed standards that include aspects of transparency should be addressed using the presented approach. For example, this may refer to important standards like ISO/IEC 42001 [13] regarding management systems for AI, or ISO/IEC 12792 [14], which directly addresses a taxonomy for transparency of AI systems. Furthermore, the development towards harmonized standards and guidelines for the implementation of the AI Act should be taken into account.

It must be noted that the considered and compared documents were released in the time span between 2019 [6] to 2024 [7]. These past five years have been characterized by an unprecedented level of disruption in AI technology, such as the development of large language models and generative AI models for image and video synthesis. These new types of models blur the line between human and AI capabilities – accompanied by a substantial shift in the perception of AI systems, their potentials, societal impacts and risks, and regulatory requirements. Hence, the documents, even though seemingly released in quick succession, must already in part be viewed in their individual "historical" context, explaining, for example that the AI Act [7] heavily addresses the challenges of GPAI, while in IEEE 7000 [11], many transparency requirements common for machine learning are absent. This explains in part the heterogeneity of the analyzed documents, such as concerning requirements with respect to GPAI models.

Further on, our analysis encountered challenges due to the inherent complexity and occasional ambiguity of the sources. The fulfillment of specific criteria was not always explicitly stated in a single, easily identifiable section, necessitating an interpretative evaluation of the

overall narrative of the document. To maintain objectivity and ensure the accuracy of our classifications, our research team implemented a systematic review process. This involved meticulous discussion and careful cross-referencing of the documents to ensure our evaluations were anchored in the text. This systematic approach, guided by the framework set out in our methodology, aimed to limit the influence of subjective judgments and provided an accurate representation of each document's stance on transparency requirements, as depicted in Tab. 3.

6 Conclusion and Outlook

In conclusion, our study offers an initial framework for evaluating and comparing ethical concepts within standards and regulations, specifically focusing on the notion of transparency in AI. Its application has highlighted that the definitions of transparency differ to a considerable degree among the considered documents. These differences can, to some extent, be attributed to the different purposes of these documents.

However, in the absence of a standardized framework for representing and comparing these definitions and their differences, the variety of notions arguably hinders the underlying goal of establishing a shared understanding across stakeholders.

Future research should expand upon this approach by assessing the criteria for transparency against a broader spectrum of references, which should include documents in the field of standardization on the one hand, e.g., regulations, standards, or guidelines; and on the other hand key academic publications as well as high-impact white papers.

In the European context, the implementation of the AI Act represents a particularly promising case for the application of our methodology as it necessitates the formulation of more specific and detailed standards for AI. By assessing how the various concepts and requirements for transparency adopted in different documents align with those found in the AI Act, our approach can help to identify key areas for creating a harmonized regulatory framework. These efforts should not only be consistent with the AI Act but also seek to elaborate more detailed standards for sector-specific applications.

Ultimately, our approach is designed to create clear and precise definitions for transparency dimensions and corresponding operationalizable criteria, representing the building blocks for agile AI governance frameworks. By doing so, we equip policymakers and industry stakeholders with fundamental tools to ensure AI trustworthiness and enhance their capacity to adapt to the ongoing developments in technology and the evolution of ethical standards in the AI field.

References

1 Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion*, 58:82–115, 2020. doi:10.1016/J.INFFUS.2019.12.012.

2 ASAM e.V., Advanced Data Controls Corp., Ansys Inc, AVL List GmbH, BTC Embedded Systems AG, DENSO Corporation, Deutsches Zentrum für Luft- und Raumfahrt e.V., Edge Case Research, e-SYNC Co. Ltd., FIVE, Foretellix Ltd, Fraunofer-Institut für Kognitive Systeme IKS, Hexagon Manufacturing Intelligence, iASYS Technology Solutions Pvt. Ltd, Institute of Communication and Computer Systems (ICCS), Oxfordshire County Council, RISE Research Institutes of Sweden, Robert Bosch GmbH, Siemens Digital Industries Software, SOLIZE Corporation, Technische Universität Braunschweig Institut für Regelungstechnik, and WMG University of Warwick. ASAM OpenODD: Concept Paper, October 2021.

3 China Academy of Information and Communications Technology (CAICT). White Paper on Trustworthy Artificial Intelligence, 2021.

4 DIN, DKE. German Standardization Roadmap on Artificial Intelligence (2nd edition), 2022.

5 Filip Karlo Došilović, Mario Brčić, and Nikica Hlupić. Explainable artificial intelligence: A survey. In *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 0210–0215, 2018.

6 European Commission / Independent High-Level Expert Group on Artificial Intelligence. Ethics Guidelines for Trustworthy AI, 2019.

7 European Parliament and the Council of the European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending certain Union legislative acts (Artificial Intelligence Act), June 2024. Official Journal of the European Union L 218, 12.7.2024. ELI: http://data.europa.eu/eli/reg/2024/1689/oj.

8 Heike Felzmann, Eduard Fosch Villaronga, Christoph Lutz, and Aurelia Tamò-Larrieux. Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns. *Big Data & Society*, 6(1):2053951719860542, 2019. doi:10.1177/2053951719860542.

9 J. Gruber. *Public Finance and Public Policy, Fifth Edition*. Worth Publishers / Macmillan Learning, 2016.

10 Benjamin Haibe-Kains, George Alexandru Adam, Ahmed Hosny, Farnoosh Khodakarami, et al. Transparency and reproducibility in artificial intelligence. *Nature*, 586(7829):E14–E16, 2020.

11 Institute of Electrical and Electronics Engineers (IEEE). IEEE 7000-2021: IEEE Standard Model Process for Addressing Ethical Concerns during System Design, 2021.

12 ISO/IEC. ISO/IEC 22989: Information technology — Artificial intelligence — Artificial intelligence concepts and terminology, April 2021.

13 ISO/IEC. ISO/IEC 42001: Information technology — Artificial intelligence — Management system, December 2023.

14 ISO/IEC. ISO/IEC DIS 12792: Information technology — Artificial intelligence — Transparency taxonomy of AI systems, July 2024.

15 ISO/IEC/IEEE. ISO/IEC/IEEE 12207:2017: Systems and software engineering — Software life cycle processes, November 2017.

16 National Institute of Standards and Technology (NIST). Artificial Intelligence Risk Management Framework (AI RMF 1.0), 2023.

17 Organisation for Economic Co-operation and Development (OECD). Scoping the oecd ai principles, 2019. doi:10.1787/d62f618a-en.

18 Organisation for Economic Co-operation and Development (OECD). OECD Framework for the Classification of AI Systems, 2022.

19 Organisation for Economic Co-operation and Development (OECD). OECD/LEGAL/0449: Recommendation of the Council on Artificial Intelligence, 2023.

20 Maximilian Poretschkin, Anna Schmitz, Maram Akila, Linara Adilova, Daniel Becker, Armin B. Cremers, Dirk Hecker, Sebastian Houben, Michael Mock, Julia Rosenzweig, Joachim Sicking, Elena Schulz, Angelika Voss, and Stefan Wrobel. *Guideline for Designing Trustworthy Artificial Intelligence: AI Assessment Catalog*, February 2023.

21 Wojciech Samek and Klaus-Robert Müller. Towards explainable artificial intelligence. *Explainable AI: interpreting, explaining and visualizing deep learning*, pages 5–22, 2019. doi:10.1007/978-3-030-28954-6_1.

22 VDE Verband der Elektrotechnik. VDE SPEC 90012: VCIO based description of systems for AI trustworthiness characterisation, April 2022.