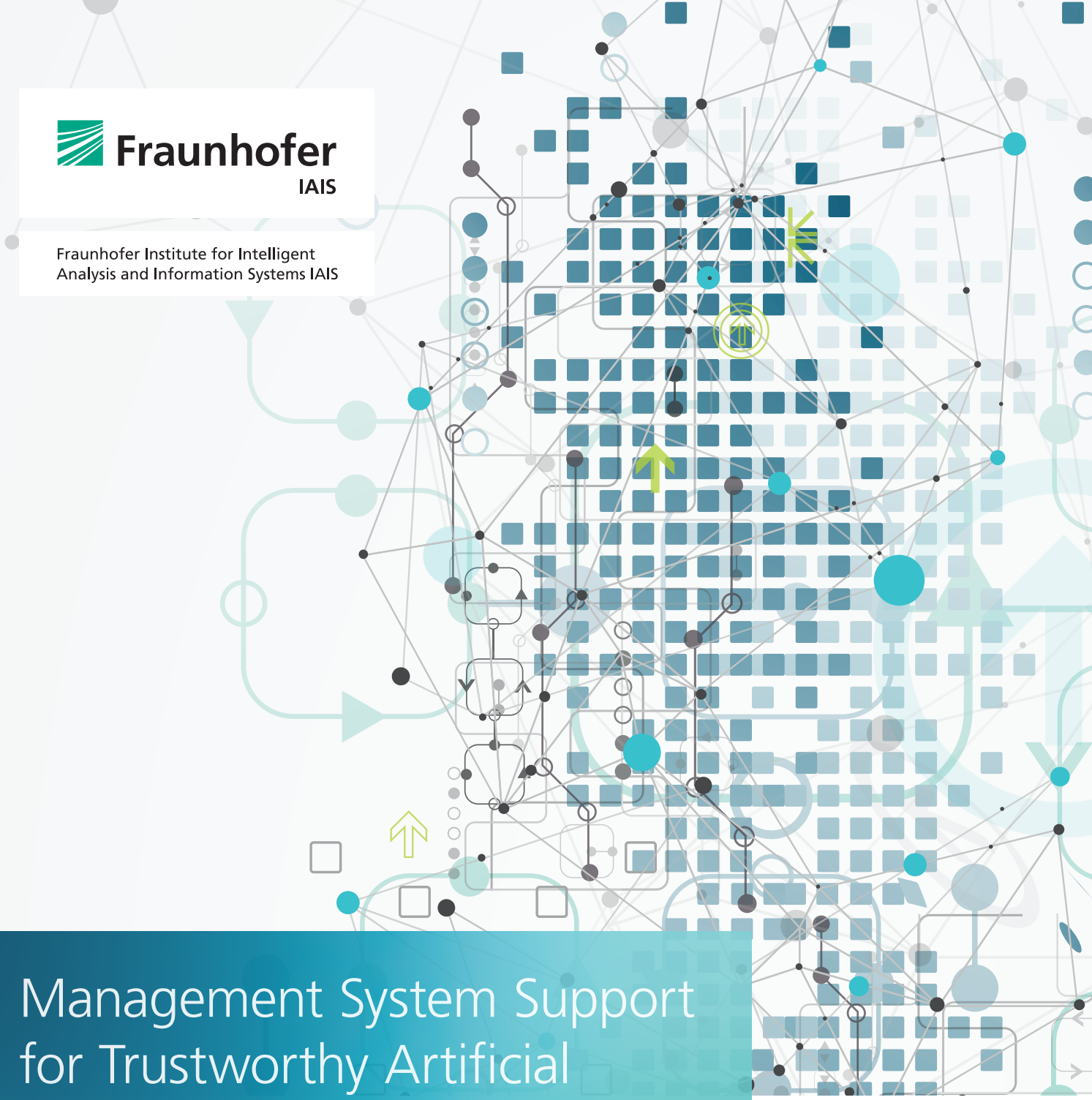

Fraunhofer IAIS

Fraunhofer Institute for Intelligent Analysis and Information Systems IAIS

Management System Support for Trustworthy Artificial Intelligence

A comparative study

www.iais.fraunhofer.de/ai-management-study

Fraunhofer IAIS

Fraunhofer Institute for Intelligent
Analysis and Information Systems IAIS

Management System Support for Trustworthy Artificial Intelligence

A comparative study

Authors
PD Dr. Michael Mock
Anna Schmitz
Linara Adilova
Dr. Daniel Becker
Prof. Dr. A.B. Cremers
Dr. Maximilian Poretschkin

October 2021

www.iais.fraunhofer.de/ai-management-study

Content

List of tables

List of abbreviations

AI	Artificial Intelligence
GDPR	General Data Protection Regulation
GRC	Governance, Risk and Compliance
ISMS	Information Security Management System
MS	Management System
MSS	Management System Standard
JTC	Joint Technical Committee
SC	Sub-committee
WD	Working Draft
WG	Working Group

Standards and other significant documents

AIC4 catalog	AI Cloud Service Compliance Criteria Catalog, published by the BSI
AIMS	International Standard for AI Management System (under development by ISO/IEC JTC 1/SC 42/WG 1)
AIMS Draft	Working Draft of the International Standard for AI Management System (ISO/IEC WD 42001, under development by ISO/IEC JTC 1/SC 42/WG 1)
ALTAI	Assessment List for Trustworthy AI, published by the High Level Expert Group on AI
ISO/IEC 17021-1:2015	International Standard for Conformity Assessment - Requirements for bodies providing audit and certification of management systems
ISO/IEC 17065:2012	International Standard for Conformity Assessment – Requirements for bodies certifying products, processes and, services
ISO/IEC 27001:2017	International Standard for Information Security Management System – requirements
ISO/IEC 23894	International Standard for AI Risk Management (under development by ISO/IEC JTC 1/SC 42)
ISO/IEC 31000:2018	International Standard for Risk Management
ISO/IEC 38500:2015	International Standard for governance of IT for the organization
ISO/IEC 9000 series	Series of International Standards for Quality Management Systems

Committees or organizations

BSI	(German) Federal Office for Information Security (German: "Bundesamt für Sicherheit in der Informationstechnik")
CASCO	ISO Committee on Conformity Assessment
DIN	German Institute for Standardization (German: "Deutsches Institut für Normung")
EC	European Commission
HLEG	High-Level Expert Group on AI
IEC	International Electrotechnical Commission
ISO	International Standardization Organization
ISO/IEC JTC 1	The ISO and IEC Joint Technical Committee 1 which deals with IT related topics
ISO/IEC JTC 1/SC 42	Sub-committee 42 of ISO/IEC JTC 1 which deals with AI
ISO/IEC JTC 1/SC 42/WG 1	Working Group 1 of ISO/IEC JTC 1/SC 42 which deals with Foundational Standards and is responsible for the development of AIMS

Foreword

The purpose of this study is to explore the role of management systems in the context of trustworthy AI.

ISO/IEC JTC 1/SC 42/WG 1, the ISO/IEC working group which deals with Foundational Standards on AI, is developing a standard for AI management systems (AIMS), that is supposed to support organizations in defining strategies, objectives and technical-organizational measures for the trustworthy use of AI systems. AIMS is in its initial stage of development and currently has the status of a working draft (AIMS Draft).This study compares the AIMS Draft against the requirements for AI of the European High Level Expert Group on AI, the proposed EU regulation on AI, and the AIC4 catalog of the German Federal Office for Information Security.

It should be noted that Fraunhofer IAIS is a member of DIN NA-043-01-42 AA – Künstliche Intelligenz, the German national mirror committee of SC 42. Moreover, it should be noted that the AIMS Draft, in its current stage, is not publicly available, however, an overview as well as an introduction to major aspects will be given in this study.

This study is sponsored by Microsoft.

Executive Summary

Artificial Intelligence (AI) technologies have a crucial impact on the economy and society and bear the potential for further relevant progress. Given the operational risks resulting from the processing of large amounts of data, on which Machine Learning (ML) technologies, in particular, are based, as well as the building of societal trust in this regard, a lot of organizations are currently working on establishing requirements or corporate goals regarding trustworthy AI. Moreover, worldwide accepted standards for the development and application of AI technologies in commercial and industrial settings are needed, as they represent an important step towards the safe and interoperable use of AI. In addition to industry, regulatory bodies have also turned their attention to AI technologies. In particular, the European Commission has recently set a milestone by publishing a proposal for AI regulation. Once in place, the regulation might have an effect comparable to the European General Data Protection Regulation that has an impact outside Europe.

For organizations using AI, the goal to be responsible, trustworthy, and compliant with (upcoming) regulation, should be significantly reflected in their governance, risk, and compliance (GRC) strategy. In general, management systems are a well-established means for structuring and managing activities within an organization to reach its goals. In particular, an appropriately designed management system can help organizations address the specific risks and new challenges inherent in AI technologies in a structured manner. In order to generate evidence of their responsibility and accountability at the management level, organizations often consult management system standards that capture worldwide accepted best practices. Currently, ISO/IEC JTC 1/SC 42/WG 1, the joint working group of the International Standardization Organization (ISO) and the International Electrotechnical Commission (IEC) that deals with foundational standards for AI, is developing an international standard for AI Management System (AIMS). AIMS is in its initial stage of development and, to date, has the status of a Working Draft[1] (AIMS Draft).

The purpose of this study is to explore the role of management systems in the context of trustworthy AI and to investigate the extent to which the AIMS Draft is suitable for supporting companies in establishing GRC strategies that allow them to develop and use AI technologies in a trustworthy way. For this purpose, we compare the AIMS Draft with several relevant documents that provide requirements and recommendations for upcoming regulations and standards[2]. More specifically, we consider:

- the recommendations for trustworthy AI from the High-Level Expert Group on AI (HLEG) in their "Assessment List for Trustworthy AI" (ALTAI),
- the requirements in the proposed regulation on AI by the European Commission (EC),
- and the criteria in the AIC4 catalog of the German Federal Office for Information Security (BSI), which extends the existing BSI C5 catalog for safe cloud-based services for AI services running in cloud environments.

A detailed summary of the recommendations and requirements of the three documents mentioned above is given in **Chapter 1.1**. It turns out, that all of them address similar technical AI-related issues, such as robustness, fairness, or transparency. Also, all of them take the special role of data as the basis for machine learning into account, from which they derive requirements on data quality, privacy protection, data governance, and risk monitoring after the deployment of AI systems. However, the terminology is not always consistent among the different documents, nor are there precise definitions everywhere. Later on, the AIMS Draft is compared against the documents mentioned above.

When establishing and evaluating the trustworthiness of AI technologies, it becomes clear that two different perspectives must be considered, which, furthermore, often interleave when it comes to controlling risks: First, the produced or marketed AI system should be of high (technical) quality and should, depending on the use case, satisfy certain product properties that, for example, contribute to the mitigation of risks. In this study, we

1 In its development process, an international standard goes through various stages: Preliminary Work Item (PWI), New Work Item Proposal (NP), Working Draft(s) (WD), Committee Draft(s) (CD), Draft International Standard (DIS), Enquiry Draft (ED), Final Draft International Standard (FDIS), International Standard (published document). For more information on the different stages, see also Chapter 2 of [ISO/IEC, 2021].

2 It should be noted that the AIMS Draft is not publicly available in its current stage and that Fraunhofer IAIS has access to the document as a member of the German national mirror committee of this standard under development. However, an overview of the content of the AIMS Draft is given in **Section 2.2** and aspects that are relevant for the comparison are presented in the tables in **Section 2.3**.

refer to the consideration of system properties as well as the consideration of risks resulting from the functioning of the AI system as the "product perspective". Similarly, requirements concerning the aspects noted above are denoted "product-related requirements". Second, the organization that develops, provides, or uses AI technologies should create an environment in which it is possible, and also ensure that the technical, as well as non-technical, requirements for the responsible use of AI are met. Therefore, roles, responsibilities, and processes within an organization need to be defined, established, and managed accordingly. In the following, we refer to the consideration of these aspects as the "organizational perspective". The interplay between product and organizational perspective, as well as challenges associated with them, are briefly sketched in **Section 1.2**. Moreover, these perspectives are taken up, in more concrete form, in **Chapter 3**, where possible approaches to certification are discussed.

Chapter 2, the main part of this study, presents a detailed comparison of the requirements and recommendations of the AIMS Draft with those from the HLEG (Assessment List for Trustworthy AI), the European Commission (Proposal for AI Regulation), and the German Federal Office for Information Security (AIC4 catalog). The chapter starts with a general introduction to management systems (see **Section 2.1**) followed by an introduction to the AIMS Draft in particular (see **Section 2.2**). For the subsequent comparison, we distinguish between requirements and recommendations regarding organizational processes (see **Section 2.3.1**) and guidance regarding technical properties of AI systems that should be considered within these organizational processes (**Section 2.3.2**).

For the technical requirements (see **Section 2.3.2**), the structure of the comparison is aligned with the framework for trustworthy AI presented in the Fraunhofer IAIS audit catalog for trustworthy AI. From the comparison, we conclude that the AIMS Draft asks organizations to establish processes that take care of the product-related requirements formulated by the HLEG, the EC, and the BSI. However, it should be noted that in the AIMS Draft technical aspects are only addressed by "controls", i.e., recommendations, that are listed in its annex. It will be up to regulatory and certification bodies to transform these controls into certification schemes – see also the related discussion in **Section 3.2.1**. All in all, we consider the AIMS Draft sufficient to cover the technical dimensions of trustworthy AI in the sense mentioned above, while leaving sufficient flexibility to organizations regarding the choice and implementation of their AI management system.

Regarding the requirements and recommendations for organizational processes (see **Section 2.3.1**), we come to a similar overall conclusion as in **Section 2.3.2**. We also see differences in terminology here, in particular concerning the notion of risk. While the proposed regulation on AI clearly sees risk in the context of safety, health, and fundamental rights of persons, the definition of risk in the AIMS Draft is more general and allows focus on potential positive or negative effects on the organization itself. Apart from following a very general definition of risk, the AIMS Draft requests organizations to carry out impact assessments that should explicitly consider risks for external stakeholders, for example, the risk that the user of an AI system could be discriminated against by a decision of the AI system. Hereby, it is left open whether the impact assessment is carried out as part of the risk assessment or separately. This is an example where we see that the proposed regulation on AI and the AIMS Draft address similar issues, but they use different terminologies (risk vs. impact) for it. As a side remark, the "impact assessment" also differentiates the AIMS Draft from other management system standards, making clear that such standards would not be sufficient to demonstrate compliance with the proposed regulation on AI. We see that the requirements and controls of the AIMS Draft cover a large part of what is demanded in the proposed regulation on AI, with the notable exception that the European Commission requires the establishment of a "Quality Management System" – see the discussion in **Chapter 4** for more details.

The four documents in the focus of this study can be the basis for certification, but certification schemes still need to be developed. Such schemes are intended to describe, among other things, by whom, under which conditions and according to which procedure a system is audited and, if successful, certified. **Chapter 3** discusses the various steps to be taken towards certification. **Section 3.2.1** sketches the way towards certification of AI management systems based on the AIMS Draft. Clearly, the AIMS Draft paves the way towards certification using the established procedures for certification of management systems that regulate the interplay between accreditation bodies, certification bodies, and applicants for certification. The controls provided by the AIMS Draft achieve a sufficient level of granularity such that, for example, certification bodies can derive suitable certification schemes. **Section 3.2.2** addresses the broader issue of achieving certification of AI systems themselves. Being developed, deployed, and monitored by a company that is governed by a certified AI management system, helps and can be the basis for, but is not sufficient for, assuring the trustworthiness of the AI system itself on the product level. For this purpose, more detailed requirements and assessment procedures on the technical level are needed. As an example, the approach described in the Fraunhofer IAIS audit catalog for trustworthy AI is briefly presented.

Chapter 4 gives an overall summary and concluding remarks. We see that, while the AIC4 catalog basically reflects the HLEG requirements adopted to the domain of cloud-based services, the AIMS Draft goes further and, interestingly, coincides with the proposed EU regulation on AI in many

respects. Both require procedures for risk management – with the side note that the AIMS Draft distinguishes between risk and impact assessment – and AI-specific post-market monitoring, thus being comprehensive for the life cycle of an AI application. Clearly, the proposed EU regulation on AI encompasses both process and product-related properties, going into more detail than the AIMS Draft regarding the latter. However, there is a big overlap regarding the process-related requirements, such that conformity with the upcoming standard ISO/IEC 42001 (AIMS) will surely support companies in complying with the upcoming regulation on AI, as well as with the AIC4 criteria catalog. Having said that it must be noted that the AIMS Draft formally does not describe a quality management system as currently being required by the proposed regulation on AI, although the AI-specific requirements described by the proposed regulation on AI on such a quality management system would be fulfilled by a management system that complies with the AIMS Draft. A further concluding remark in **Chapter 4** emphasizes the importance of the EU General Data Protection Regulation (GDPR) in the context of AI. Since the European Commission aims to harmonize regulations in the further implementation of the upcoming regulation on AI, conformity with data protection requirements will play an important role in the certification of AI – this should be made very explicit in upcoming management system standards.

1. Introduction to the Trustworthiness of AI Systems

AI is a key technology that has a crucial impact on economic and societal progress. By supporting medical diagnoses, decisions on loan approval, and, in prospect, autonomous driving, AI plays an increasingly important role in our daily lives and enters more and more sensitive domains. However, apart from its huge potential, AI yields new risks. Being a data-driven technology, an AI system could, for instance, take over data-inherent discrimination. Furthermore, the functionality of AI systems could be manipulated by poisoned training data. Also, the fact that many AI methods are highly non-transparent for humans poses a challenge, for example regarding the verification of what an AI system has actually learned. For AI systems to be trustworthy, it thus must be ensured that such AI-related risks are under control. Only when trust in AI is established can its full economic and societal potential be realized.

A proven method to build trust in AI is to demonstrate conformity with recognized standards, especially by certification. The need for standardization of AI was identified and thoroughly discussed in course of the project "German Standardization Roadmap on AI" [DIN e.V. & DKE, 2020], which was led by the German Institute for Standardization (DIN). The authors of the roadmap point out that existing standards for classic IT applications like the ISO/IEC 27000 series do not cover the risks inherent to AI technologies sufficiently. Further, they state the need for certification schemes to put the certification of AI into practice.

Besides, there has been a long, interdisciplinary debate about the definition of appropriate trustworthiness requirements that reflect AI-specific risks in particular. European boards and authorities have also addressed this issue in recent years. In 2019, the HLEG published its "Ethics Guidelines for Trustworthy AI" [HLEG, 2019b] and, recently, the European Commission (EC) proposed a regulation laying down harmonized rules on AI [EC, 2021]. On a national level, the German Federal Office for Information Security (BSI) published the "AI Cloud Service Compliance Criteria Catalogue (AIC4)" [BSI, 2021], which lists requirements for the security of AI-based cloud services. **Chapter 1.1** gives an overview of the requirements which are formulated by the HLEG, the EC, and the BSI and which provide a potential basis for standardization and certification.

When it comes to implementing and verifying the trustworthiness of AI systems and their operation, it becomes clear that two different but often interleaving perspectives must be considered: product and organizational perspectives. Trustworthy AI from a product perspective requires that the AI system is of high technical quality and that it satisfies certain product properties that contribute to the mitigation of risks. However, the practical implementation of desired system properties, as well as their technical verification, are challenging, especially since the concrete requirements strongly depend on the application context as well as the specific AI technology. On the other hand, in order to monitor and maintain the technical features of an AI system and to ensure that relevant risks are permanently under control, appropriate structures and activities are needed from an organizational perspective. The importance and interplay of product and organizational perspectives in realizing trustworthy AI are briefly sketched in **Section 2.1**.

1.1 Requirements for trustworthiness by institutions and authorities

The question as to which criteria should be used to assess whether the use of AI can be considered trustworthy and how our societal values can be implemented in this future technology has been the subject of intensive research and broad societal debates[3]. The challenge in defining trustworthiness requirements is to capture the risks inherent in AI, the measurability of which is not obvious and to formulate them in a way that their fulfillment ensures those risks are under control. A lot of companies and NGOs have established their own guidelines or corporate goals for trustworthy AI[4]. However, the concrete (quantitative) criteria for a specific AI system depend on its application context and the implemented AI technology, so the requirements are usually kept at an abstract level and need to be further operationalized.

3 For an overview of ethical, legal, and societal challenges see [Floridi, 2016], [Floridi, 2018]. For a review and research questions from a data-driven perspective see [Thiebes, 2020].

4 For an overview of AI ethics guidelines by companies, organizations, and states, see [Jobin, 2019].

In recent years, European boards and authorities, in particular, have taken a close look at the problem of protecting the safety, health, and fundamental rights of persons when using AI. In 2019, the High-Level Expert Group on AI (HLEG) published its "Ethics Guidelines for Trustworthy AI" [HLEG, 2019b] on behalf of the European Commission (EC). The Ethics Guidelines, which formulate seven key requirements for trustworthy AI, are a major reference point in the debate on trustworthy AI. Subsequently, the European Commission proposed a regulation laying down harmonized rules on AI [EC, 2021]. It builds on the guidelines by the HLEG and follows a risk-based approach. In particular, the proposed regulation demands that 'high-risk' AI systems undergo a conformity assessment before they can enter the European market, where it is up to the standardization organizations to provide technical specifications and to define detailed technical requirements and measures by which conformity can be reached.

Apart from European bodies, national bodies have also taken the first steps towards formulating initial requirements for the development, deployment, and use of AI. In Germany, the Federal Office for Information Security (BSI) has published the AI Cloud Service Compliance Criteria Catalogue (AIC4) [BSI, 2021] which lists requirements for the security of AI-based cloud services.

The following subsections give an overview of the requirements, which are formulated by the HLEG, the EC, and the BSI in the documents mentioned and which provide a potential basis for standardization and certification.

1.1.1 HLEG requirements

In June 2018, the European Commission set up a group of experts to provide advice on its AI Strategy. This High-Level Expert Group on Artificial Intelligence (HLEG) comprised 52 experts, bringing together representatives from academia, civil society, and industry. In April 2019, this independent expert group published 'Ethics Guidelines on trustworthy AI' [HLEG, 2019b], followed by 'Policy and Investment Recommendations for Trustworthy AI' [HLEG, 2019a]. The overall work of the HLEG has been central for European policymaking initiatives and, in particular, for upcoming legislative steps.

In its Ethics Guidelines, the HLEG introduces four ethical principles for trustworthy AI which are: "respect for human autonomy", "prevention of harm", "fairness" and "explicability". From these, they derive the following seven key requirements for trustworthy AI: "human agency and oversight", "technical robustness and safety", "privacy and data governance", "transparency", "diversity", "non-discrimination and fairness", "environmental and societal well-being" and "accountability." In this regard, the HLEG highlights that the requirements depend on the specific application and that they need to be implemented throughout an AI system's life cycle. It is noteworthy that the key requirements are formulated in a way that leaves wide scope for interpretation because the HLEG does not specify to what extent or by which particular measures those requirements may be fulfilled. Furthermore, trade-offs between the key requirements should be taken into account.

The Ethics Guidelines contain a preliminary list for self-assessment which is meant as guidance for achieving trustworthy AI that meets the seven key requirements. This assessment list was finalized following a period of stakeholder consultation comprising, amongst others, fifty interviews with companies. The final 'Assessment List for Trustworthy Artificial Intelligence' [HLEG, 2020] (ALTAI) was published in July 2020 and the presentation of the key requirements in this chapter is oriented to the structure of ALTAI. Here it should be noted that the HLEG requests that companies conduct a fundamental rights impact assessment before their self-assessment regarding the seven key requirements.

1.1.1.1 Human agency and oversight

"AI systems should support human autonomy and decision-making, as prescribed by the *principle of respect for human autonomy*. This requires that AI systems should both act as enablers to a democratic, flourishing, and equitable society by supporting the user's agency and foster fundamental rights, and allow for human oversight." [HLEG, 2019b]

Human agency and autonomy: In the first place, it is requested that the use of AI is disclosed to persons who interact with an AI system or who use its results. Moreover, the risk that humans become over-reliant, disproportionately attached, or addicted to an AI system should be mitigated, and the possibility that users could be affected in a way that their behavior or decisions are illegitimately or maliciously manipulated should be avoided.

Human oversight: It is requested that AI systems can be overseen by humans as appropriate. Human oversight should be supported by the AI system itself, for instance by a stop button or detection mechanisms. Further, it is requested that humans who take over oversight are adequately prepared, for example by specific training.

1.1.1.2 Technical robustness and safety

"A crucial component of achieving Trustworthy AI is technical robustness, which is closely linked to the *principle of prevention of harm*. Technical robustness requires that AI systems be developed with a preventative approach to risks and in a manner such that they reliably behave as intended, while minimizing unintentional and unexpected harm, and

preventing unacceptable harm. This should also apply to potential changes in their operating environment or the presence of other agents (human and artificial) that may interact with the system in an adversarial manner. In addition, the physical and mental integrity of humans should be ensured." [HLEG, 2019b]

Resilience to Attack and Security: In the first place, it is requested that threats regarding technical and security-related issues and their potential impact are understood. Moreover, an AI system should be compliant with relevant cybersecurity standards, and it should also be resilient against AI-specific attacks and vulnerabilities such as, for instance, data poisoning. The system should be tested accordingly, and the end-users should be informed about the given security coverage as well as updates.

General Safety: It is requested that risks related to fault or misuse of an AI system are continuously assessed and monitored, for instance by evaluating adequate risk metrics. Furthermore, fault tolerance should be ensured. If necessary, end-users should be informed, and a review of the system's technical robustness and safety may be initiated.

Accuracy: It should be ensured that an AI system operates at a sufficient level of accuracy, which should also be monitored and communicated to the user. Additionally, the data used in the context of the AI system should be relevant, representative, and of high quality.

Reliability, Fall-back plans, and Reproducibility: The reliability of an AI system should be continuously evaluated with appropriate verification and validation methods. Further, relevant data should be documented and, if appropriate, specific contexts/scenarios should be taken into account to ensure reproducibility. Moreover, fail-safe fallback plans, as well as procedures for handling low confidence of results and the risks of continual learning, should be established.

1.1.1.3 Privacy and data governance

"Closely linked to the *principle of prevention of harm* is privacy, a fundamental right particularly affected by AI systems. Prevention of harm to privacy also necessitates adequate data governance that covers the quality and integrity of the data used, its relevance in light of the domain in which the AI systems will be deployed, its access protocols, and the capability to process data in a manner that protects privacy." [HLEG, 2019b]

Privacy: The impact of the AI system on the rights to privacy and data protection should be assessed. Furthermore, there should be the possibility to flag issues concerning those rights.

Data Governance: Regarding personal data, the HLEG mainly refers to the requirements contained in the GDPR. They should be implemented by technical measures ('privacy-by-design) as well as by oversight mechanisms for data processing. Moreover, the AI system should be compliant with relevant standards for data management and governance.

1.1.1.4 Transparency

"This requirement is closely linked with the *principle of explicability* and encompasses transparency of elements relevant to an AI system: the data, the system, and the business models." [HLEG, 2019b]

Traceability: There should be mechanisms and procedures for record-keeping to allow for traceability of the AI system's decisions or recommendations. Also, the quality of input and output data should be monitored.

Explainability: The technical processes, as well as the reasoning behind an AI system's decisions, should be explained to users and affected persons to the degree possible and appropriate. Especially, it should be checked whether the given explanation is actually understandable and informative to the persons addressed.

Communication: Users should be informed about the purpose, capabilities, and limitations of an AI system in an appropriate manner. For instance, they should be informed about its level of accuracy. Moreover, they should be provided material on how to adequately use the system.

1.1.1.5 Diversity, non-discrimination, and fairness

"In order to achieve Trustworthy AI, we must enable inclusion and diversity throughout the entire AI system's life cycle. Besides the consideration and involvement of all affected stakeholders throughout the process, this also entails ensuring equal access through inclusive design processes as well as equal treatment. This requirement is closely linked with *the principle of fairness*." [HLEG, 2019b]

Avoidance of Unfair Bias: The creation or reinforcement of unfair bias in the AI system should be avoided. For this, an appropriate definition of fairness should be chosen in consultation with impacted groups. Tests and monitoring should ensure that the data is diverse and representative of the specific target group. Moreover, the developers should be made aware of potential biases, and they should take adequate measures in the algorithm design. Furthermore, there should be the possibility to flag issues concerning discrimination or unfair bias.

Accessibility and Universal Design: An AI system should address the widest possible range of users and, in particular, be accessible to them. Therefore, universal design principles should be considered during the planning and development of an AI system, especially taking into account and, if possible, involving, potential end-users with special needs. It should be ensured that no group of people is disproportionately affected by the results of the AI system.

Stakeholder Participation: Stakeholders should participate in the design and development of an AI system. Also, they should be consulted after deployment; to give feedback, for instance.

1.1.1.6 Societal and environmental well-being

"In line with the *principles of fairness* and *prevention of harm*, the broader society, other sentient beings, and the environment should be also considered as stakeholders throughout the AI system's life cycle. Sustainability and ecological responsibility of AI systems should be encouraged, and research should be fostered into AI solutions addressing areas of global concern, such as the Sustainable Development Goals. Ideally, AI systems should be used to benefit all human beings, including future generations." [HLEG, 2019b]

Environmental Well-being: The environmental impact of an AI system should be assessed throughout its life cycle and its entire supply chain. Measures should be taken to reduce this impact.

Impact on Work and Skills: It is requested that the impact of an AI system on human work and skills is understood. Moreover, impacted workers should be consulted and informed about the operation and capabilities of an AI system. The risk of de-skilling the workforce should be tackled, and workers should be provided with adequate training opportunities in case they require new skills with regard to the AI system.

Impact on Society at large or Democracy: The impact of an AI system on democracy and society at large, i.e., beyond individual users, should be assessed. Measures should be taken to ensure that an AI system cannot cause harm to society and democratic processes.

1.1.1.7 Accountability

"The requirement of accountability complements the above requirements and is closely linked to the *principle of fairness*. It necessitates that mechanisms be put in place to ensure responsibility and accountability for AI systems and their outcomes, both before and after their development, deployment, and use." [HLEG, 2019b]

Auditability: Internal or external audits of an AI system should be facilitated as appropriate, for instance by documentation of relevant processes and record-keeping.

Risk Management: Organizations are requested to have a risk management system in place which should comprise, as appropriate, oversight by a third-party or an AI ethics review board over the implemented ethical and accountability practices. Further, adherence to the key requirements should be continuously monitored and, if necessary, trade-offs between key requirements should be discussed and the respective decisions explained. Further, it should be ensured that the persons involved in the risk management process are adequately trained, especially regarding the applicable legal framework. Finally, there should be the possibility for third parties to report issues and risks related to the AI system and to get access to appropriate redress mechanisms.

1.1.2 European proposal for the regulation of AI

In April 2021, the European Commission published a "Proposal for Regulation of the European Parliament and of the Council laying down harmonized rules on AI (Artificial Intelligence Act) and amending certain union legislative acts" [EC, 2021]. With its proposal, the European Commission is reacting to calls by the European Council, especially, for a review of the existing relevant legislation concerning the challenges raised by AI ([EC, 2021], p. 3). Further, the proposal follows several AI-related resolutions by the European Parliament ([EC, 2021], p. 3) and is a result of extensive consultation with stakeholders and experts ([EC, 2021], p. 8-9). One of the main objectives of the proposed regulatory framework is to "ensure that AI systems placed and used on the Union market are safe and respect existing law on fundamental rights and Union values" ([EC, 2021], p. 4).

The proposed AI regulation follows a risk-based approach and, besides prohibiting certain AI practices, specifies technical requirements for 'high-risk AI systems' in the Union market. Apart from product-related requirements, it also formulates management-related obligations for relevant parties in the context of 'high-risk AI systems', especially for providers. Moreover, the document prescribes how compliance with certain requirements in the regulation is to be assessed and declared. In particular, a conformity assessment procedure is required for high-risk AI systems before their placing on the Union market, and harmonized standards and common specifications may be consulted.

1.1.2.1 Risk-based approach

The proposal lays down harmonized rules following a risk-based approach. Here, it distinguishes between "prohibited AI practices", "high-risk AI systems", and others. This categorization of AI systems is predominantly oriented towards their potential negative impact on the health, safety, or fundamental rights of persons.

Title II of the document prohibits certain (subliminal) AI techniques that could materially distort the behavior of persons, as well as AI systems for social scoring that could lead to the unjustified or disproportionate detrimental treatment of persons and – with exceptions – AI-based real-time biometric identification systems in publicly accessible spaces. However, AI systems that are exclusively developed and used for military purposes are not in the scope of the proposed regulation ([EC, 2021], Article 2).

The first chapter of Title II of the proposed regulation gives a definition of "high-risk" AI systems. Hereby, it distinguishes between two types. On the one hand, it considers AI systems which are a safety component of a product, or which are a product themselves, which is covered by existing Union harmonization legislation[5]. Within these, such AI systems are categorized as high-risk, if the product they are embedded in, or if the AI system itself as a product, is required to undergo a third-party conformity assessment with a view to the placing on the market or the putting into service of that product in accordance with the Union harmonization legislation ([EC, 2021], Article 6). Regarding the New Legislative Framework, these are, in particular, such products as "machinery, toys, lifts, equipment, and protective systems intended for use in potentially explosive atmospheres, radio equipment, pressure equipment, recreational craft equipment, cableway installations, appliances burning gaseous fuels, medical devices, and in vitro diagnostic medical devices" ([EC, 2021], p. 26).

On the other hand, the definition considers stand-alone AI systems, i.e., systems that are not necessarily a safety component of a product or a product themselves. Annex III of the proposed AI regulation specifies pre-defined areas where such systems are likely to pose a high risk of harm to the health and safety or the fundamental rights of persons and which are therefore classified as 'high-risk'. These application areas include, amongst others, AI systems for biometric identification, AI systems for management and operation of critical infrastructure, AI systems for determining access or assigning persons to educational and vocational training institutions, AI systems used in employment, workers' management and access to self-employment, e.g., for the recruitment and selection of persons or for monitoring or evaluation of persons in work-related contractual relationships, AI systems that decide on access to certain essential private and public services and benefits, in particular AI systems used to evaluate credit scores or to establish priority in the dispatching of emergency first response services, AI systems to detect the emotional state of natural persons or to detect 'deep fakes' for the evaluation of the reliability of evidence in criminal proceedings, AI systems used in migration, asylum and border control management and AI systems intended to assist judicial authorities in researching, interpreting and applying the law.[6]

1.1.2.2 Requirements for high-risk AI systems

Chapters 2 and 3 of Title III of the proposed AI regulation formulate mandatory requirements for high-risk AI systems, their providers, and other relevant parties. These requirements comprise technical aspects on the one hand and procedural or management-related obligations on the other. Especially, the requirements concern the whole lifecycle of an AI system.

Technical/System-related requirements
The technical or system-related requirements are described in Chapter 2 of Title III and cover aspects of data quality, record-keeping, transparency, human oversight, reliability, and cybersecurity.

The requirements regarding data quality specify that the training, validation, and test data shall be "relevant, representative, free of errors and complete", and that they shall take into account characteristics of the specific application area. Furthermore, the data shall be examined with a view to possible biases. If necessary, for example, if personal data is processed, security or privacy-preserving measures shall be taken. ([EC, 2021], Article 10)

Moreover, it is required that high-risk AI systems keep a record of data or events to an extent that enables appropriate tracing as well as monitoring of its functioning. In this regard, the document formulates minimum requirements for biometric identification systems. However, for a specification of such logging capabilities, the proposal refers to harmonized standards and common specifications. ([EC, 2021], Article 12)

Further, it is demanded that the operation and functioning of the AI system are made transparent to users to the extent required for reasonable use. Therefore, the AI system shall be

5 A full list of that legislation, based on the New Legislative Framework or on other Union legislation, is given in Annex II of [EC, 2021].

6 For a complete description, see Annex III of [EC, 2021].

accompanied by instructions for use, which, amongst others, shall provide information about the system's performance and its limitations, about circumstances or potential misuse that may lead to risks, as well as about necessary maintenance measures, measures for human oversight, and, especially, technical measures which support users in interpreting the output. ([EC, 2021], Article 13)

Additionally, it is required that the use of AI in some application contexts is disclosed, regardless of whether the AI system is 'high-risk' or not. This applies, in particular, to contexts where natural persons interact with AI systems, where natural persons are exposed to emotion recognition or biometric identification systems, and for AI systems that generate so-called "deep fakes". However, exceptions are made for certain AI systems that are particularly authorized by law to detect or prevent criminal offenses. ([EC, 2021], Article 53)

Besides this, it is required that human oversight by natural persons is enabled and effectively implemented. Here, human oversight has the goal of minimizing or preventing relevant risks during the use of the AI system. According to the proposal, it comprises the monitoring of the system's operation with the opportunity to detect and address dysfunction or unexpected performance, as well as the ability to disregard or override the output and to intervene in or interrupt the operation. Human oversight measures can be implemented by the design of an AI system, for instance with human-machine interface tools, or they can be implemented by the user. The document does not contain any concrete mandatory measures in this regard; however, it highlights that persons assigned for human oversight shall have a thorough understanding of the AI system, be aware of risks such as automation bias, and, if available, use tools to understand and interpret the output correctly. ([EC, 2021], Article 14)

Finally, it is demanded that high-risk AI systems "achieve, in the light of their intended purpose, an appropriate level of accuracy, robustness, and cybersecurity, and perform consistently in those respects throughout their lifecycle". While the document does not further specify the "appropriate level", it points out risks that, if applicable, shall be addressed by technical solutions. Regarding a consistent performance, the proposal requires on the one hand that, if the AI system continues to learn after deployment, the risk of biased outputs, due to so-called feedback loops, are duly addressed. On the other hand, the system shall be robust against foreseeable errors and inconsistencies within its environment, for which technical solutions like backup or fail-safe plans may be considered. Lastly, cybersecurity shall be ensured with a particular focus on the integrity and availability of the AI system. Here, the proposal explicitly names "data poisoning" and "adversarial examples" as AI-specific vulnerabilities/attacks that shall be addressed by technical solutions. ([EC, 2021], Article 15)

Management-related obligations of providers

Apart from (technical) system-related requirements, Chapter 3 of Title III also places obligations on providers of high-risk AI systems with respect to their organization and management.

One of the main obligations of providers, as defined by the proposed regulatory framework, is that they shall "put a quality management system in place that ensures compliance with this [the proposed] Regulation", and thus in particular with the previously described system-related requirements. Amongst others, the quality management system shall include "techniques, procedures and systematic actions for design, design verification, development, (…) quality control, quality assurance" as well as "examination, test and validation procedures", "technical specifications, including standards, to be applied", "systems and procedures for data management (and) record-keeping", "procedures related to the reporting of serious incidents", "resource management", and an "accountability framework" ([EC, 2021], Article 17).

In addition to the aforementioned techniques, procedures, and sub-systems within the quality management system, two further requirements for the quality management system are particularly emphasized: a risk management system and a post-market monitoring system shall be in place. Concerning the risk management system, it is required that risks related to the AI system are identified, analyzed, and evaluated in a continuous iterative process throughout the lifecycle of the AI system, and that appropriate mitigation and control measures are adopted. Moreover, the AI system shall be tested before being placed on the market to ensure that its performance is appropriate and consistent for the intended purpose and that all other system-related requirements are fulfilled ([EC, 2021], Article 9).

Further, the provider is required to have a post-market monitoring system in place that can collect and analyze relevant data and thus enable the provider to evaluate continuing compliance with the system-related requirements described in Chapter 2 of Title III. The post-market monitoring system shall be based on a post-market monitoring plan about which the document states that 'the Commission shall adopt an implementing act laying down detailed provisions establishing a template for the post-market monitoring plan and the list of elements to be included in the plan'. ([EC, 2021], Article 61)

1.1.2.3 Obligations regarding assessment and declaration of conformity

The proposed regulatory framework requires that high-risk AI systems undergo a conformity assessment procedure ([EC, 2021], Article 19) before they are placed on the market or put into service. While it depends on the type of AI system

whether the conformity assessment is conducted by a third party or based on internal control, the main subject of the assessment is a technical documentation of the AI system in any case. If compliance with the system-related requirements, as well as regarding the risk management system, has been successfully demonstrated, the provider shall formally declare conformity with the regulation ([EC, 2021], Article 48) and especially affix the CE marking ([EC, 2021], Article 49).

According to the proposal, a provider of a high-risk AI system is obliged to draw up technical documentation "in such a way to demonstrate that the high-risk AI system complies with the requirements set out in (Chapter 2 of Title III[7]) and provide national competent authorities and notified bodies with all the necessary information to assess the compliance of the AI system with those requirements" ([EC, 2021], Article 11). Annex IV specifies which aspects the technical documentation shall contain at the least.

Amongst others, a general description of the AI system shall be drawn up together with a detailed technical description comprising "the process for its development", "information about the data used", "validation and testing procedures", "test reports", "design specifications", "metrics used to measure (…) compliance", "technical solutions adopted to ensure continuous compliance", an assessment of the measures for human oversight and of the measures to facilitate the interpretation of the output, and a discussion of trade-offs. Moreover, the documentation shall contain "detailed information about the monitoring, functioning and control of the AI system", a "description of the risk management system" and the post-market monitoring plan, and a "list of the harmonized standards applied" ([EC, 2021], Annex IV).

In this regard, the European Commission points out that standards and other technical specifications play a key role in providing "the precise technical solutions to achieve compliance with [the] requirements" set out in Chapter 2 of Title III ([EC, 2021], p. 13) and that "common normative standards for all high-risk AI systems should be established." ([EC, 2021], p. 20). In particular, the regulatory framework provides that conformity with the requirements in the proposal shall be equivalent to conformity with certain harmonized standards or common specifications, to the extent that they cover the requirements. ([EC, 2021], Articles 40 and 41)

The proposal envisages two types of conformity assessment procedures for high-risk AI systems ([EC, 2021], Article 43). On the one hand, for most stand-alone systems, as listed in Annex III of the proposal, it requires a conformity assessment based on internal control. Here, the provider verifies whether its quality management system is in conformity with Article 17 of the proposal, and examines the technical documentation, which he/she then uses as the basis for assessing compliance of the AI system with the requirements described in Chapter 2 of Title III ([EC, 2021], Annex VI). On the other hand, high-risk AI systems which fall under the New Legislative Framework ([EC, 2021], Annex II, Section A) shall follow the third-party conformity assessment as required under the respective legal act. However, the conformity assessment shall be extended in the manner that the notified body (third-party) examines the technical documentation and, if necessary, conducts further tests to assess the conformity of the AI system with the requirements set out in Chapter 2 of Title III. ([EC, 2021], Annex VII)

In more complex business scenarios which also involve importers, distributors, and manufacturers, some of the provider's obligations regarding the conformity assessment or the CE marking may be transferred to other parties or may have to be fulfilled by them as well. For these cases, we refer to Chapter 3 of Title III of the proposed AI regulation.

1.1.3 AIC4 catalog

On a national level, the German Federal Office for Information Security (BSI) has taken the first steps to promote the security of AI systems and services. Having identified the gap that AI-specific security threats are not covered by established IT security standards, the BSI has taken the initiative to tackle this gap, in the first place with a focus on AI services running in cloud environments. In February 2021, the AI Cloud Service Compliance Criteria Catalogue (AIC4) [BSI, 2021] was published as an extension to the internationally recognized BSI Cloud Computing Compliance Criteria Catalogue (C5) [BSI, 2020a]. The AIC4 catalog formulates AI-specific requirements across the AI lifecycle that are related to the use of AI. With the criteria catalog, the BSI sets a baseline level of security for AI-based cloud services and supports users in evaluating the trustworthiness of such services.

The requirements of the catalog are structured according to the following eight criteria areas: "security and robustness", "performance and functionality", "reliability", "data quality", "data management", "explainability" and "bias". These areas represent the security areas of AI cloud services according to the BSI. Further, the BSI states that the criteria in the catalog together make up the minimum requirements for professional AI usage. They are supposed to ensure that AI service providers use state-of-the-art processes for the development, testing,

7 Chapter 2 of Title III contains the technical system-related requirements, as well as the requirements regarding the risk management system and the technical documentation.

validation, deployment, and monitoring of AI. In addition to each criterion, the catalog gives supplementary information which indicates how the criterion can be achieved.

The AIC4 catalog is explicitly restricted to AI-specific requirements. The preliminary criteria of the AIC4 catalog refer to the C5 catalog [BSI, 2020a] for general requirements concerning cloud computing and, in particular, compliance with the C5 catalog is a prerequisite for compliance with AIC4.

1.1.3.1 Security and robustness

The criteria in this area are:
- Continuous Assessment of Security Threats and Countermeasures
- Risk Exposure Assessment
- Regular Risk Exposure Assessment
- Testing Learning Pipeline Robustness
- Testing of Model Robustness
- Implementation of Countermeasures
- Residual Risk Mitigation

The area "security and robustness" deals with security threats like attacks, malfunction, or misuse, which especially evolve if the confidentiality or integrity of data is violated. The criteria stipulate that such threat scenarios are monitored and evaluated on a regular basis. In particular, security threats to the learning process of the model as well as to the deployed model shall be tested. Therefore, specifically designed attacks shall be used which, for instance, are based on manipulated training or input data. Moreover, it is required that appropriate countermeasures are taken in the design and learning process of the model as well as during deployment. Especially, the confidentiality and integrity of the data need to be ensured to mitigate and/or prevent the risks related to security threats.

1.1.3.2 Performance and functionality

The criteria in this area are:
- Definition of Performance Requirements
- Monitoring of Performance
- Fulfillment of Contractual Agreement of Performance Requirements
- Model Selection and Suitability
- Model Training and Validation
- Business Testing
- Continuous Improvement of Model Performance
- Additional Considerations when using Automated Machine Learning
- Impact of Automated Decision-making
- Regular Service Review

The criteria in "performance and functionality" aim to ensure that the AI service has sufficient performance and that deviations are handled appropriately. The requirements relate to the full lifecycle. In the first place, it is demanded that performance goals are defined which are appropriate for the application context given. The design and choice of the model shall be suitable for the given tasks and the provider needs to be aware of their limits. Further, it is required that the training and evaluation of the model follow recognized methodologies. In particular, if identified during validation, inaccuracies like under-/overfitting of the model shall be addressed. Moreover, there needs to be a process for monitoring the performance and reviewing the service during operation based on user feedback and failure reports. If necessary, a re-training or changes to the system should be initiated. Apart from the performance-related criteria, it is also required that humans have the opportunity to update or modify decisions by the AI service.

1.1.3.3 Reliability

The criteria in this area are:
- Resource Planning for Development
- Logging of Model Requests
- Monitoring of Model Requests
- Corrective Measures to the Output
- Handling of AI-specific Security Incidents
- Backup and Disaster Recovery

The area 'reliability' contains criteria for the smooth operation of the AI service. Apart from ensuring that all necessary resources are provided, potential security incidents during operation shall be prevented or handled appropriately if they occur. One of the main requirements in this area is that relevant data is logged and, where applicable, monitored. For instance, irregularities in the interactions with the service must be detected in order to prevent or back-track failures and other security incidents. For the same reasons, a roles and rights concept is required, which shall make sure that only authorized people can overwrite the outputs of the service. If incidents occur, back-ups of data and model shall be available to enable fast recovery of the service. Furthermore, security incidents should be consolidated into new threat scenarios, which are considered in the risk assessment process.

1.1.3.4 Data quality

The criteria in this area are:
- Data Quality Requirements for Development
- Data Quality Requirements for Operation
- Data Quality Assessment
- Data Selection

- Data Annotation
- Preparation of Training, Validation, and Test Data

The area 'data quality' deals with the quality of data used by the AI service. On the one hand, the provider needs to define data quality requirements for development and, if applicable, requirements to ensure the accuracy of annotations. These requirements shall be used to assess data in their selection process. Moreover, it must be ensured that the distribution and split of training, test, and validation data are appropriate for the application context. On the other hand, the provider shall define data quality requirements for operation. Especially, if users can provide data to the AI service, these requirements must be transparent. The quality of data shall be checked regularly during operation and, if necessary, corrective measures shall be taken.

1.1.3.5 Data management

The criteria in this area are:
- Data Management Framework
- Data Access Management
- Traceability of Data Sources
- Credibility of Data Sources

The criteria in 'data management' aim at a structured and secure provision of data. It is demanded that the provider has a data management framework in place that gives guidance and/or establishes processes for the acquisition, distribution, storage, and processing of data. In particular, this framework relates to the full lifecycle and shall manage the data used for the development, operation, and further improvement of the AI service. Moreover, it is required that the provider documents the origin of data and assesses its credibility. Credibility is especially related to the confidentiality and integrity of data, which shall be ensured by measures like access authorization and encryption.

1.1.3.6 Explainability

The criteria in this area are:
- Assessment of the required Degree of Explainability
- Testing the Explainability of the Service

The area 'explainability' requires that decisions made by the service are explained to the extent that is necessary and appropriate. The need for explainability of the AI service shall be analyzed, particularly taking into account its application context and criticality. Moreover, appropriate transparency methods shall be used to provide explanations. Thereby, trade-offs between explainability and performance, as well as limitations of the transparency methods, must be addressed and communicated in the system description.

1.1.3.7 Bias

The criteria in this area are:
- Conceptual Assessment of Bias
- Assessing the Level of Bias
- Mitigation of detected Bias
- Continuous Bias Assessment

The area 'bias' shall ensure that bias does not have critical implications on the functionality and security of the AI service. The provider is required to assess the possibility of bias and its potential implications and to test the service and data with respect to critical bias on a regular basis. If existent, the provider is required to take appropriate measures to mitigate intolerable bias. Further, threats or limitations regarding the mitigation of bias shall be made transparent to users.

1.2 Challenges regarding the implementation and testing of requirements for trustworthy AI

In order to realize the full societal and economic potential of AI, it is important that AI systems are developed and operated according to high quality standards and that trust is established by making this visible. A proven method of achieving this is to demonstrate conformity with broadly accepted standards, but many of these do not yet exist for AI technologies. An overview of relevant trustworthiness requirements that can be brought to life through standards was given in **Section 1.1** However, in order to determine measurable criteria for a given AI system the guard rails presented in **Section 1.1** must first be concretized for classes of use cases. In addition, when implementing trustworthiness requirements, two often mutually dependent perspectives must be taken: product and organizational perspectives. These in turn are associated with different kinds of challenges. On the one hand, challenges arise in terms of implementing system properties for trustworthy AI and generating technical evidence for it. On the other, there is a high effort of an organizational nature, e.g., with regard to the definition and implementation of test procedures, processes to ensure data quality, or the preparation of documentation.

From a product perspective, trustworthiness requires that the AI system is of high (technical) quality and that it satisfies certain product properties which contribute to the mitigation of risks. However, the practical implementation of desired system properties, as well as the generation of technical evidence for it, are often challenging, especially because the concrete requirements strongly depend on the specific application context and because there is no obvious method of measurement for most properties. An additional challenge arises from the fact that there is often a chain of distributed responsibilities for the quality of an AI system. Since existing IT testing procedures are not readily transferable to AI technologies, methods for

testing and verification of AI systems are an active research area. Section 1.2.1 elaborates on challenges from a product perspective and gives a broad overview of the state-of-the-art research on testing methods for AI. This section may be skipped if the technical challenges are already known.

What is not immediately apparent from the product perspective is that, from an organizational perspective, it takes a great deal of effort to constantly ensure that risks are under control and to guarantee traceability in this regard. For example, test procedures for the AI system as well as processes to ensure data quality or the preparation of documentation must be defined and continuously operated. Thus, to realize the trustworthy use of AI within an organization, there is a particular need for appropriate structures, clearly defined responsibilities, and roles, as well as non-technical measures and processes, especially for risk management and human oversight. Section 1.2.2 illustrates the importance of the organizational perspective and its interaction with the product perspective on trustworthiness.

1.2.1 Product perspective

Due to the complexity of AI systems, various challenges arise regarding the realization and verification of their quality. Especially for AI systems based on machine learning, novel risks arise from the processing of data. How these risks are to be evaluated usually depends strongly on the application context. Because AI specifics are not adequately addressed in existing IT testing procedures, assessment schemes and testing tools for AI are the subjects of active research.

As can be seen from the overview of the requirements in Section 1.1, trustworthiness in the context of AI has many facets. AI applications are often based on machine learning (ML) techniques that learn patterns in so-called training data and build a model to apply what has been learned to unknown data (but structurally comparable to the training data). Due to the central importance of the training data for the functionality of ML-based AI systems, new challenges arise in implementing trustworthy AI. One difficulty regarding reliability, for example, is to specify the application domain of the system as precisely as possible and to cover it accordingly with training data in order to counteract malfunction or even systematic model weaknesses. In open-world contexts in particular, it is usually not possible to quantify the application domain precisely. In addition, training data poses novel security risks such as data poisoning, where the integrity of the system is violated by deliberate manipulation of the database. Apart from the areas of reliability and security, the realization of transparency and fairness in AI systems is also discussed. The functioning of the underlying models is often difficult to understand, even for experts, due to the large number of parameters. Human interpretability and methods for explaining AI results are the subjects of active research. Similarly, various disciplines are researching concepts and technical methods for implementing fairness in AI, since AI systems, as data-driven technologies, tend to incorporate data-inherent discrimination into their models. Last but not least, another key challenge is that the learning process of AI systems continues in principle even during operation, so that appropriate monitoring and control mechanisms are to be developed to detect and prevent the learning of incorrect behavior.

It may also be the case that trade-offs need to be made between the mitigation of different types of risks when creating trustworthy AI. For example, an increase in performance, such as the recognition performance in object recognition on image data by deep neural networks, can be at the expense of interpretability. Another example is that an increase in transparency, by disclosing all hyperparameters of a model, for example, can lead to new attack vectors in terms of IT security.

Another challenge in ensuring the quality of AI systems arises from the fact that their development is distributed along a value chain that is very different from the development of conventional software. In the following, as AI systems we denote complex IT systems that include machine learning-based components for performing particular tasks. Since ML models are often specified over millions (sometimes billions) of parameters, AI systems rely, in particular, on the processing of large amounts of data, for which corresponding IT infrastructures and computing power are required. As a result, organizations that develop or use AI systems often rely on third-party components. On the one hand, such components can be purchased as an AI product, which means that it is produced by an external provider but is deployed in the internal infrastructure (hardware) of the organization without the provider having further access to it. On the other hand, there is the possibility of purchasing AI components as a service which means that an AI component is made available for use but is still deployed in the infrastructure of its provider. Cloud service providers play an important role here, providing the necessary computing capacity, infrastructure, and corresponding basic AI services such as optical character recognition, video analysis, speech-to-text conversion, text-to-speech conversion, translation, text analysis, or intelligent search. In order to simplify the use of such services and their immediate adaptation to user needs, AI services are usually encapsulated in a simple graphical user interface or calls to libraries in the respective programming language. Organizations purchasing AI services for their system can therefore save time, effort, and resources for local development. However, incorporating third-party components into an AI system often leads to the fact that these have a significant influence on the quality of the AI system, without the organizations receiving comprehensive information or insight from the providers. Moreover, cloud-based services require additional consideration when

it comes to the trustworthiness of the AI system, especially concerning privacy and security (see [Chiregi, 2018]).

Just as with the practical implementation of trustworthiness, a key challenge with the technical verifiability is that both the concrete requirements and the way they are realized strongly depend on the specific application context. One thrust in the area of AI testing and verification is to define concrete metrics and qualitative requirements that measure the performance of an AI system with respect to a specific target variable and relate this to a specific requirement for trustworthy AI (see [Verma, 2018], [Papineni, 2002], [Salimans, 2016], [Weng, 2018], [Hess, 2018]). There are several approaches to evaluate an AI system with respect to these metrics. One approach is to develop structured questionnaires (see [Gebru, 2018], [Arnold, 2019], [Mitchell, 2019], [Madaio, 2020]), which evaluate the extent to which a given AI system meets qualitative or quantitative criteria for trustworthiness. The second approach is to develop testing tools that measure the quality of AI systems (or of their building blocks, such as datasets) in a (partially) automated way (see [Bellamy, 2019], [Saleiro, 2018], [Nicolae, 2018], [Arya, 2020], [Santos, 2017], [Nori, 2019]). The challenge with these approaches, however, is to establish criteria that are proportionate for a given application context, given the plethora of AI technologies and their diverse uses. The metrics used, and, in particular, the respective target values, are often specific to a particular class of use cases, which makes it difficult to compare results. As can be seen from the requirements and recommendations for trustworthy AI summarized in **Section 1.1**, the HLEG, the EC, and the BSI do not follow the approach of specifying concrete metrics and target values. Their requirements and recommendations are kept on a rather abstract level and need to be further operationalized and, if possible, quantified, for the specific use case.

A proven approach to operationalize use case-specific trustworthiness requirements is a risk-based testing approach. Risk-based approaches are found, for example, in the classic concepts of IT Security[8] and Functional Safety, where the requirements for resistance to manipulation or unintentional misconduct can lead to very different technical requirements for different systems. In particular, risk-based testing approaches are intended to ensure comparability of the test results of different systems, despite very different individual requirements. Since, however, existing methods for risk-based testing do not cover AI specifics appropriately and are thus not readily transferable to AI systems, the underlying concept of risk-based testing is an important object of research in the area of trustworthy AI. Currently, many research activities are focused on transferring concepts of Functional Safety to AI systems, where the use case of autonomous driving plays a major role (see [Huang, 2017], [Burton, 2017]). An example of a framework that is considered in the field of safety is 'Claims, Arguments, and Evidence'. It was used for AI by [Zhao, 2020]. In this framework, claims serve as system requirements, evidence provides information that is related to the system considered, and arguments are the way evidence is linked to a claim.

Other research activities explore the questions of how to conduct conformity assessments for AI and how to verify the trustworthiness of AI systems in the context of an audit (see [Hallensleben, 2020]). As mentioned before, the integration of third-party components (as product or service) into AI systems poses a challenge when it comes to evaluating system properties. This is because external components yield additional trustworthiness requirements and often no deep insight into these components is possible. However, professional cloud customers may, acting in the same way as auditors, want to conduct their own risk assessment and employ additional security measures for which they require transparent and detailed information about each external component. For example, a framework for evaluating the transparency and fairness of AI services is proposed by [Antunes, 2018]. Here, transparency includes various sub-dimensions such as awareness, explanations and interpretability of results, and access to the documentation and the component itself. The latter is a key prerequisite for the proper assessment of an AI system. Especially in the case of AI systems that include several external ML products and services that may even be provided from different countries, it is likely that different standards need to interact when auditing the system.

Another complexity that arises from the use of cloud-based AI services, and also from internal machine learning modules that learn continuously, is that updates or modifications of their functioning are made in small intervals of time. Thus, a simple, one-time assessment or audit of the AI system might be not appropriate. Accordingly, there is a general call for continuous monitoring of such AI systems, but no clear view has yet emerged on the duration of validity of test results or the intervals and criteria for the reassessment of AI systems.

1.2.2 Organizational perspective

Organizations must make appropriate preparations to meet and demonstrate compliance with system-level requirements for trustworthy AI. On the one hand, many system-related requirements cannot be implemented by technical solutions

8 See, for example, BSI-Grundschutz or ISO/IEC 27001:2013 for specifications for an information security management system or the Common Criteria [CC 3.1, 2017] for a methodology for testing IT products.

alone, but their fulfillment requires human involvement or at least human oversight. On the other hand, preparing and/or conducting (internal or external) audits of an AI system takes a great deal of effort for organizations that, for instance, need to draw up respective technical documentations. To manage these new efforts, and also to control the distributed responsibilities for the quality of an AI system, organizations face the challenge of establishing appropriate roles, structures, and processes.

Judging by the current state of the EU-level documents and the AIC4 catalog as presented in **Section 1.1**, fulfilling requirements for trustworthy AI will entail a considerable amount of work and effort from an organizational perspective. As briefly discussed in **Chapter 1.2.1**, the research regarding technical trustworthiness properties of AI systems, especially their measurement and implementation techniques, is broad and ongoing. However, it is already apparent that effective measures to ensure system properties like robustness, fairness, or transparency often require access to the training data, the design process, and the representation of the output. This provides an intuition that the trustworthiness of AI systems as required by the European Commission for example can, in general, only be guaranteed by a broad analysis of both the system and its environment, accompanied by careful risk mitigation measures and checkups. Before organizations address this challenge, they must create internal awareness of the responsibilities and new tasks it entails. AI-related governance and policies within an organization should, in particular, take into account the responsible development of AI systems. So far, there is still no general agreement on how trustworthy AI development is to be achieved from an organizational perspective. As an example, [Askell, 2019], who also emphasizes that organizational policies play a vital role for the development of trustworthy AI, gives the following definition of responsible AI development:

"Responsible AI development involves taking steps to ensure that AI systems have an acceptably low risk of harming their users or society and, ideally, increase their likelihood of being socially beneficial. This involves testing the safety and security of systems during development, evaluating the potential social impact of the systems before release, being willing to abandon research projects that fail to meet a high bar of safety, and being willing to delay the release of a system until it has been established that it does not pose a risk to consumers or the public."

Moreover, they even discuss that organizations may cooperate to avoid competition that results in pressure to invest less effort and money than necessary to maintain trustworthiness.

The above description of responsible development, similar to the requirements summarized in **Section 1.1**, needs to be further operationalized in order to give concrete guidance for a specific organization. Furthermore, ensuring the trustworthiness of an AI system does not end at the development stage: post-market monitoring and many levels of support and maintenance are needed throughout the entire system lifecycle. Here, too, the opportunities of both technical design and AI management should be taken into account, as pointed out by [Mittelstadt, 2019] and [Brundage, 2020]. In particular, many features that technically constitute a trustworthy AI system intrinsically require process regulation. For example, to ensure the reliability of an AI system that continuously learns during deployment, it is not sufficient that it fulfills certain technical requirements by design. For keeping risks under control, e.g., the risk that the AI system learns a wrong behavior, whether due to data drift or to manipulated training data, its performance should be monitored. In addition, risks due to potential changes to the system environment should be assessed regularly. Many of these issues cannot, or rather should not, be addressed by technical solutions alone but should at least involve human oversight and the option for human intervention in critical situations. Thus, human action is clearly supportive and, in critical application contexts, even necessary for identifying, limiting, and managing adverse effects or damages, direct or mediated, that an AI system may have or inflict on humans, its environment, or even the organization providing or using it – especially when the system is not aligned with system-level requirements. To achieve appropriate human activities such as oversight, corresponding structures, roles, and procedures need to be established.

While internal procedures for testing and risk assessment may be established to generally ensure the quality of an AI system or for competitive advantages, AI systems may also undergo tests and assessments to demonstrate conformity with obligatory requirements as appropriate. For this purpose, too, organizations must coordinate corresponding preparations. As indicated in the proposed regulation for AI, providers of 'high-risk' AI systems will be required to draw up technical documentation, which can be used as the basis for assessing conformity with the technical requirements in the regulation. Resources and competence will be needed to deal with legal or compliance issues and correspondence with notified bodies if necessary.

Another challenge facing organizations is the issue of scattered responsibilities. As already touched upon in **Section 1.2.1**, in the case of ML technologies, many (third) parties along the value chain of AI development and operation often have a stake in the quality of an AI system. These are parties that, for example, generate or collect data, provide the necessary IT infrastructure for processing large amounts of data, or provide ready AI products or services. Apart from responsibilities being distributed to third parties, also the different steps of development, deployment, and use, which are often performed by the

provider or user of the AI system itself, entail a corresponding chain of actors that all have an impact on the quality of the AI system. As [Coeckelbergh, 2020] highlights, the question of assigning responsibilities is important and all actors involved should be considered in this regard. The importance of clearly defining responsibilities is also shown by [Peters, 2020] who considers the chain of responsibilities with an eye toward errors and failures of the AI system that can potentially cause harm.

A concrete example of a chain of responsibilities oriented along the development steps of an AI system up to deployment is given by [Zweig, 2018]:

1. Algorithm design and implementation include a lot of important decisions concerning the ML model. If the model is used in a ready-made form, [Zweig, 2018] emphasizes that all parameters selected in the process should be known.
2. Selection of methods for the problem-solution requires knowledge of the application area to identify and evaluate all possible risks.
3. Data collection and selection yield multiple requirements that must be taken care of, including privacy and fairness in particular.
4. Construction of the system requires the proper and adequate estimate of its performance.
5. Embedding in the social process. Here, transparency about the system and the interpretability of the results are important.
6. Re-evaluation of the system after deployment
7. Liability detection

Another view is taken by [Zicari, 2021] who proposes that responsible parties should be described in terms of "ecosystems" of particular parts of an AI system rather than seeing responsibilities as being distributed along a chain.

[Toreini, 2019] sees a similar issue as that of scattered responsibilities in the concept of trust. He proposes that a distinction be made between trust, as an essential foundation for social contracts, and ethics, as one of several aspects affecting this trust. Further, he claims that trust as a social concept mainly relates to the organization that provides an AI system and not to technical details. As such, [Toreini, 2019] sees trust being distributed along the "trust chain" within (and between) organizations and, in general, all agents that are involved in AI production and development. Being transparent about their AI systems and their functionalities, and enabling stakeholders to participate in the development process and operations are two examples of measures that organizations can take to foster trust. As for the other organizational challenges depicted in this section, it becomes clear that appropriate structures, roles, processes, and activities within organizations are needed to realize trustworthy AI.

Summarizing, we can see that both the product and the organizational perspective have to be considered when aiming at the trustworthy development and use of AI systems. The challenges involved become even harder when taking into account the rapid evolution of AI techniques and novel business models based on AI. In the next chapter, we will investigate the role of AI management systems to support companies dealing with these challenges.

2. Analysis of the AI Management System Standard Draft ISO/IEC WD 42001

The last chapter has illustrated that many technical, as well as organizational, challenges are associated with trustworthy AI. The implementation and assurance of technical requirements must be coordinated and, to permanently ensure that relevant AI risks are under control, appropriate organizational procedures need to be established. Therefore, resources, roles, and responsibilities also need to be organized and managed accordingly. Comparable management challenges exist in the field of IT security. Here, the ISO/IEC 27001 international standard specifies requirements for an information security management system (ISMS), that successfully supports organizations dealing with these security-related challenges. Hence, one possible way to address AI-related challenges within an organization can be an AI-specific management system.

Management systems are an established means within an organization to systematically support the execution of purposeful and accountable management. Their aim is the setting up of policies and processes in order to reach organizational objectives and, thus, affect different parts of an organization from governance to technical-organizational measures. In particular, management systems build a strong foundation for a framework for governance, risk, and compliance (GRC). Accordingly, management systems have been standardized for over 30 years in order to support organizations to generate evidence of their responsibility and accountability. Among the most popular systems is ISO 9001, the international standard for quality management systems (for which over 1 million certificates ([Lambert, 2017], p. 37-40) have been issued), and ISO/IEC 27001, the international standard for information security management. An introduction to management systems (MS) and management system standards (MSS) in general is given in **Section 2.1**.

In view of the increasing importance and spread of artificial intelligence applications worldwide, ISO and IEC are developing a set of standards to address different aspects of the use of AI technologies. A particular one that is currently under development by ISO/IEC JTC 1/SC 42/WG 1 is the international standard for AI management systems (AIMS). Currently, in the stage of a Working Draft (AIMS Draft), it defines requirements and controls with regard to AI management which are intended to help organizations deal with the specific issues and risks that arise from the use of AI and to reach their AI related objectives in view of these challenges. **Section 2.2** presents the AIMS Draft and gives an overview of the most relevant aspects.

The main part of this chapter, **Section 2.3**, analyzes to what extent AIMS, in its current version, is suitable for supporting providers or users of AI systems in meeting relevant trustworthiness requirements. For this purpose, the AIMS Draft is compared with requirements and recommendations formulated by the High-Level Expert Group on AI, the European Commission, and the German Federal Office for Information Security, as described in **Section 1.1**. The analysis distinguishes between technical requirements and higher-level requirements (process-view) that address processes and structures within an organization.

2.1 Overview of management systems in general

Management systems are a suitable tool for organizations to address the challenges and risks in achieving their goals in a structured and responsible manner. ISO defines management systems as

"(…) the way in which an organization manages the interrelated parts of its business in order to achieve its objectives. These objectives can relate to a number of different topics, including product or service quality, operational efficiency, environmental performance, health and safety in the workplace and many more." [ISO, 2021]

[Kersten, 2020] explains a management system for a topic X to be

"generally anything that is used to identify the key objectives for topic X, achieve those objectives, and monitor their maintenance" (translated from German).

Thus, management systems concern the governance layer of an organization as well as its management and technical-organizational measures implemented at a lower level. In general, they help to combine all organizational units and processes that are directed to set objectives into a clear framework. Depending on the kind of objectives, its scope can range from covering the whole organization to managing

a particular sector or function of the organization. Typical parts or tasks within a management system include

- "formulating objectives in the form of policies,
- analyzing risks and opportunities for these objectives,
- defining roles or responsibilities for specific (sub-) objectives,
- (…) planning and implementing processes and the measures required to achieve them,
- and planning, implementing and evaluating reviews of the achievement of objectives" (translated from German, see [Kersten, 2020]).

This description gives a sense that a management system is an appropriate tool for an organization to address the challenges regarding trustworthy AI described in **Section 1.2.2**. More generally, management systems are a suitable tool for setting up a reliable framework for governance, risk, and compliance (GRC) across an organization.

Standardization – being a proven method to establish interoperability, common terms, and definitions, as well as a basis for trust in technical systems and processes – is thus also carried out for management systems (MSs). ISO describes the purpose of standardization of MSs in the following way:

- "ISO management system standards (MSSs) help organizations improve their performance by specifying repeatable steps that organizations consciously implement to achieve their goals and objectives, and to create an organizational culture that reflexively engages in a continuous cycle of self-evaluation, correction and improvement of operations and processes through heightened employee awareness and management leadership and commitment." [ISO, 2021]

The first MSS in the world, ISO 9001, international standard for a quality management system (QMS), was introduced in 1987 [Lambert, 2017]. Today, there are more than 60 MSSs by ISO and IEC. Apart from QMS which is the most certified ISO MSS to date, also ISO/IEC 27001, international standard for information security management (ISMS), is among the most popular ones.

A lot of MSSs are broadly recognized and trusted to the extent that in some sectors it has become a convention to have, for example, a QMS in place. Many customers or business partners see certification against a QMS as a benchmark for ensuring and continuously improving the quality of products or services. In addition to meeting customer requirements or ensuring the quality of products or services (in the case of QMS), MSSs can also help organizations comply with regulatory requirements. Furthermore, having a certified MS in place generates evidence of the responsibility and accountability of an organization and can ease compliance dependencies in particular. In certain cases, certification against an MSS is even a regulatory requirement in itself. For example, a QMS is prescribed for many companies in the healthcare sector. Moreover, the proposed regulation on AI by the European Commission also mentions a QMS for providers of high-risk AI systems as a requirement and, in addition, points out the importance of (harmonized) norms and standards for the (technical) elaboration or concretization of requirements[9].

A harmonized high-level structure (HLS) was set up to increase the scalability and interoperability of management system standards.[10] "Developed by ISO, the HLS provides identical structure, text and common terms and definitions for all future ISO MSSs. Now, all ISO's management systems standards could be aligned, facilitating full integration of several standards into one management system in a single organization" [Lambert, 2017]. According to the HLS, the main part of an MSS is to be built in the following way:

1. Scope
2. Normative references
3. Terms and definitions
4. Context of the organization
5. Leadership
6. Planning
7. Support
8. Operation
9. Performance evaluation
10. Improvement

Relevant aspects of these ten structural elements are described below.

Regarding the scope of a management system standard (point 1), Annex SL of the ISO/IEC Directives provides a number of definitions. In addition to defining a management system as a

"set of interrelated or interacting elements of an organization to establish policies and objectives, as well as processes to achieve those objectives

- Note 1 to entry: A management system can address a single discipline or several disciplines.
- Note 2 to entry: The management system elements include the organization's structure, roles and responsibilities, planning and operation."

9 See p.13, p.20 and p.32 of [EC, 2021].

10 The high-level structure as well as identical sub-clause titles, identical text, common terms, and core definitions for management system standards are defined in Annex SL of [ISO/IEC, 2021].

It also introduces the notions of "generic MSS" and "sector-specific MSS" and further distinguishes between so-called "type A" and "type B" MSSs. Type A management system standards provide requirements that characterize the different stages of an MS (points 4 to 10 of the HLS). Thus, type A MSSs are predestined for certification. However, the requirements provided in a type A MSS do not prescribe the practical way in which management processes or other technical-organizational measures must be implemented, as this is highly dependent on the context and the characteristics of the respective organization. Instead, recommendations about the implementation are usually given as controls in the annex, for instance, ISO/IEC 27001:2015 contains 114 controls, or in type B MSSs [ISO/IEC, 2015b]. In general, type B MSSs provide guidelines or more specific requirements, possibly related to a particular topic, that guide organizations in the implementation of their MS. Type B MSSs are, in turn, not certifiable. Management standards are another kind of standard that may provide implementation guidance for particular aspects of an MS, for example, ISO 31000:2018 [ISO, 2018], the international standard for risk management.

Due to the harmonization of their high-level structure, many ISO MSSs contain uniform terms and definitions (point 3). Some of them, which are listed in most MSSs under point 3, are given in the following:

objective: result to be achieved
- Note 1 to entry: An objective can be strategic, tactical, or operational.
- Note 2 to entry: Objectives can relate to different disciplines (such as finance, health and safety, and environment). They can be, for example, organization-wide or specific to a project, product, or process. (…)
- Note 4 to entry: In the context of [topic] management systems, [topic] objectives are set by the organization, consistent with the [topic] policy, to achieve specific results.

policy: intentions and direction of an organization as formally expressed by its top management

process: set of interrelated or interacting activities that uses or transforms inputs to deliver a result
- Note 1 to entry: Whether the result of a process is called an output, a product or a service depends on the context of the reference.

requirement: need or expectation that is stated, generally implied or obligatory
- Note 1 to entry: "Generally implied" means that it is custom or common practice for the organization and interested parties that the need or expectation under consideration is implied.
- Note 2 to entry: A specified requirement is one that is stated, e.g., in documented information".[11]

As already outlined, points 4 to 10 of the HLS characterize the different stages of an MS, whether in terms of requirements (type A MSSs) or guidelines (type B MSSs). An important feature of MSSs is that they are not value-based, but instead they support organizations in setting up a management system based on their organizational context (point 4). This includes, amongst others, the identification of all interested parties and stakeholder expectations. According to MSSs, organizations should determine the scope and functionality of their respective MS only after they understand the specific challenges, risks, and expectations within their internal and external environment.

The introduction of a management system also requires commitment and directional decisions by leadership (point 5). In particular, leadership shall ensure that the objectives to be achieved by the management system are established. Moreover, they shall be aligned with the overall governance of the organization and supported by respective policies and guidelines. The HLS also sets out requirements on how to formulate objectives, which, for example, should be as measurable as possible.[12] A reliable commitment to objectives also requires that roles and responsibilities are clearly assigned in this regard. Given the scattered responsibilities discussed in **Section 1.2.2**, this is a particularly critical aspect when it comes to the management of AI-related issues.

To achieve set objectives, appropriate procedures, processes, and technical-organizational measures need to be planned and implemented (points 6 and 8). Regarding ISMS, for example, a core objective is to protect information as an asset from possible sources of damage like attacks, errors, vulnerabilities, and nature. Typical processes that organizations establish within an ISMS are risk assessment, constant monitoring of risks, evaluation of the effectiveness of processes in place, and updating of processes and measures in order to improve them. Possible technical-organizational measures within an ISMS are, for instance, the training of employees to create awareness of security risks, access control, encryption of data, physical protection of machines and devices from accidents or attacks, video surveillance, and screen locks.

MSSs also point to the fact that appropriate support (point 7) is needed to implement the planned processes and measures. Depending on the purpose of the MS, this may range from

11 Annex SL Appendix 2 of [ISO/IEC, 2021].

12 Annex SL Appendix 2 of [ISO/IEC, 2021].

(competent) human resources and awareness to technical assets and precautions. Another important aspect of support within every MSS is documentation of the MS and other relevant information. More detailed requirements regarding documented information are given in the HLS. On the one hand, documentation of relevant information can support transparency and accountability. On the other hand, regarding the use of AI, documentation can be useful for evaluating an AI system and identifying sources of errors or failures.

Finally, a crucial aspect of the HLS is the regular evaluation of the management system in place (point 9). To this end, the organization shall perform internal audits at planned intervals within which conformity of the implemented and maintained processes with the policies and objectives established by its governance is checked. According to ISO terminology,

"Audit: systematic and independent process for obtaining evidence and evaluating it objectively to determine the extent to which the audit criteria are fulfilled".[13]

Moreover, management reviews shall be drawn up at planned intervals. The goal of performance evaluation is to ensure that deviations from requirements are detected and addressed so that the management system is continuously improved (point 10).

2.2 Introduction of the AIMS Draft

The technical committee ISO/IEC JTC 1/SC 42 Artificial Intelligence is developing a number of standards in the field of AI, addressing various aspects from general process related issues such as AI risk management to technical product-oriented standardization of issues such as ISO/IEC WD 42001, Working Draft on "Information Technology – Artificial Intelligence – Management System".

The Artificial Intelligence Management System (AIMS) standard is being developed to help organizations that provide or use AI to manage their policies and processes to achieve their objectives. The standard does not distinguish the type, size, and nature of organizations that are using it.

The main part of the AIMS Draft defines requirements regarding policies, general processes, and functions.

1. The organization is supposed to identify all the stakeholders in the context of an AI system. In particular, the roles of provider, developer, user, and partner (for example, data provider) are important. The needs and requirements of these parties are identified, and the scope of the AIMS is thus defined.
2. The top management is supposed to be dedicated to defining goals and objectives for the AIMS and ensuring that all the required resources are assigned for performing the selected policies. The policy definition is important in the context of AIMS: it must be documented, communicated, available, and require the purposes of an AI system.
3. As a consequence of having defined requirements, the organization shall identify the risks related to them. There must be a risk assessment process to identify, analyze, and evaluate risks related to AI systems, and this must be clearly documented. For each identified risk, the organization shall find a corresponding control to mitigate this risk.
4. Particularly important for an AI system provider is to identify the objectives for the system and usage of the system, that are aligned with the organizational policy and stakeholders' requirements. Examples of the objectives are given in Annex B of the draft and include fairness, security, safety, privacy, robustness, transparency and explainability, accountability, availability, maintainability, availability, and quality of training data, AI expertise.
5. AIMS shall be supported by an organization on the level of required resources, competent employees, and awareness among employees, including communicating the objectives and policies.
6. The thorough documentation requirements on the AIMS.
7. Finally, as a common requirement for a management system, the constant internal and external audit is required to check the conformity of the functioning system to the documentation and standards. The results of an audit are used for continuous improvement of the AIMS.

Annex A of the AIMS Draft provides the set of controls that can be adapted by an organization for achieving selected objectives. This list identifies main groups, and, in case of need, the organization can introduce different specific controls. The central specification in the context of a management system is a policy. On the high level, the main requirement for a policy in AIMS is to be aligned with the business strategy and values of an organization and also with the regulations and risk environment of the organization. The policy should identify the scope of AI in the organization, the objectives, and the handling of processes. It is important to mention, that AIMS-related policies might intersect with other management systems, for example, related to security and safety. In such cases, the other management systems should be updated to include AI-related controls. Implementation of an AIMS in an organization requires designating responsibility roles, for example, for risk assessments, impact assessments, asset and resource management, security, privacy, development, performance, human oversight, supplier relationship. The second aspect specified is the establishment of a related reporting system.

13 Annex SL Appendix 2 of [ISO/IEC, 2021].

The next recommendation is related to the resources needed for the AI system. These include data, human resources, and technical tools (algorithm types, data conditioning tools, optimization methods, evaluation methods, machine learning tasks, machine learning approaches). All the resources for the AI system must be documented in order to support the further impact and risk assessment. For data, it is particularly important to specify the origin, category, and classes; the algorithms must be specified in how they are preprocessing data, what machine learning approaches are used; system (computational) resources must be carefully selected and described with respect to the stage of the development; and human resources have to include data scientists, researchers in ethics and society, human oversight experts, AI researchers, and other specialists. Each of the resource types should be carefully identified with respect to the requirements of the corresponding stage of an AI system lifecycle.

An impact assessment is important to the organization of a trustworthy AI provider. The corresponding processes in the AIMS must define the evaluation of both the benefits and risks that are presented to the stakeholders, including the organization itself. This contrasts with the usual risk assessment within a management system, extending the risks with the overall impact of an AI system on all the stakeholders. The activity concerned with the analysis of impacts should be integrated into the processes ensuring the trustworthiness of AI, for example, risk management, in the organization and be properly documented for further usage. Three large groups of potentially affected subjects are individuals, society as a whole, and the organization itself. On the individuals' level, it is important to understand the risks related to privacy, interpretability, fairness, and safety. On the level of the society, the harm to the environment, the health of the society, and the alignment to culture and values shall be considered.

The next part in the AIMS Draft is the management of the development lifecycle of the AI system. The processes for responsible AI development shall be set up so that they are aligned to the overall policy and objectives with respect to the AI system. The organization has to define what constitutes the objectives of the responsible development in each exact situation, which again refers to the fairness, security, etc. (Annex B of the AIMS Draft). Each phase of the AI system lifecycle shall be managed correspondingly. The objectives of the development and success indicators are to be documented; the design solutions (such as machine learning approach, algorithms, how the model will be trained and with what data, evaluation and refinement of the models, hardware and software, code) are to be documented; strict verification and validation criteria should be fixed; deployment has to consider the possible changes in the environment and include a set of requirements that should be met every time before the new setup; most importantly, the operation shall be logged and controlled, all the updates and fixes provided correspondingly and controlled to meet the requirements of all the stakeholders involved. In particular, the last stage refers to the post-market monitoring of the AI system, where the minimal considerations include monitoring for general errors, performance expectations correspondence; monitoring of performance in case of continuous learning; monitoring for the concept drift and possible retraining; processes for updates introduction in case of failures; support processes for users. Finally, the data quality shall be controlled and documented, which is expanded into a separate control recommendation. It is important to emphasize the role of the data in any AI system. In particular, high requirements should be met by the privacy and security consideration of the data question, as well as transparency of data provenance processes. The details of the data selection shall be determined and documented, including aspects like the amount of data needed, sources of the data, types of the data. The quality and the pre-processing of the datasets are additional controls that should be defined.

The responsible use of an AI system is considered in terms of separate policy controls. In the case where an organization is using an AI system provided (developed) by another organization, the responsible-use process must be set up by defining the objectives that must be achieved (from the list of fairness, security, etc., see Annex B of the AIMS Draft). The third-party relationship shall be clearly documented to ensure disentanglement of the chain of responsibilities (as discussed in **Chapter 1**) and to ensure that all the providers are implementing the required controls for the responsible AI development/usage.

Finally, the security system be adapted to the AI system and take into account AI-specific threats (such as, for example, adversarial attacks or privacy attacks).

2.3 Analysis of the AIMS Draft regarding the trustworthiness requirements

In this section we will analyze to what extent the AI management system (AIMS) Draft proposed by ISO/IEC JTC 1/SC 42/WG 1 is suitable for supporting providers or users of AI applications in meeting relevant trustworthiness requirements. For this purpose, the AIMS Draft is compared with requirements and recommendations formulated by the High-Level Expert Group on AI (HLEG), the European Commission (EC), and the German Federal Office for Information Security (BSI), as described in **Chapter 1**.

The comparison first focuses on process-/management-related aspects. In particular, **Section 2.3.1** analyzes to what extent the AIMS Draft covers processes and procedures associated with risk management, the planning of resources, data governance, and accountability. **Section 2.3.2** follows a scheme as presented in the Fraunhofer IAIS audit catalog for trustworthy AI. It is structured along six technical dimensions of trustworthiness, supplemented by the discussion of trade-offs between those dimensions and the issue of monitoring.

2.3.1 Process-view

The requirements and recommendations as formulated by the HLEG, EC, BSI, and ISO/IEC JTC 1/SC 42/WG 1 have the objective of ensuring the responsible and trustworthy development, deployment, and use of AI systems. Major aspects of this objective are that organizations that provide or use AI systems control relevant risks on the one hand and, on the other, take responsibility for these risks. While the mitigation and control of risks can be addressed in large part through technical requirements for an AI system, none of the four instances consider system-specific requirements or recommendations in isolation from structures or processes within an organization. In particular, they emphasize the importance of regular reassessments of potential threats and risks in face of the dynamics of AI systems. Apart from risk management, the planning of resources, accountability, and data governance with a view to the deployment and use of AI systems are particularly emphasized. Only in an organizational culture that is sensitized to risks and clearly defines roles and responsibilities with regard to AI systems can a trustworthy provision or use of AI, which also requires the implementation of technical measures for risk mitigation, be ensured.

2.3.1.1 Risk management

Definition

Risk management is a complex undertaking for organizations. As described in ISO/IEC 31000:2018, the international standard for risk management, it comprises iterative processes and procedures which assist an organization to deal with factors and influences that make achieving their goals uncertain. Risk management is an iterative process comprising the assessment, treatment, monitoring, review, recording, and reporting of risk. In particular, risk management is an integral part of decision-making and contributes to the improvement of management systems.[14]

Since AI systems often tackle complex problems in uncertain environments, they do not provide automation in the sense that their actions and outcomes would be clearly predictable. Instead, the functioning of an AI system is usually opaque to humans, and its underlying decision rules evolve dynamically, so that AI gives rise to new imponderables for providers and users. Therefore, the HLEG, EC, BSI, and ISO/IEC JTC 1/SC 42/WG 1 all prominently address the issue of risk management within their requirements and recommendations.

It should be noted that the four documents considered follow different notions of risk. In general, risk describes a potential (unexpected) effect which can arise from uncertain conditions. While the HLEG and the EC predominantly consider risk in light of "negative consequences for individuals or the society" ([EC, 2021], p.2), putting particular emphasis on the safety, health, and fundamental rights of persons ([EC, 2021], p.4), ISO defines risk in the international standard for risk management as the "effect of uncertainty on objectives" (ISO/IEC 31000:2018 see [ISO/IEC, 2018]). Thus, ISO has a more neutral view on risk in the sense that effects can be considered as positive or negative. However, in the standard for risk management, risk is considered in terms of its effect or consequences for the organization itself, i.e., its objectives, while the recommendations and requirements by the HLEG and the EC aim at controlling risks with a view to human well-being. Again differently, the BSI uses the term risk in its AIC4 catalog only in the context of concrete security incidents like leakage or corruption of data or the model, failure, attack, and bias, without specifying whether their effects should be considered with respect to the organization or other stakeholders. Lastly, compared with the BSI, the HLEG, and the EC, the most general definition of risk is given in the AIMS Draft as "effect of uncertainty", deviating from the original definition in the international standard for risk management. This also leaves open whether effects on the organization itself, stakeholders, or third parties are considered here.

Comparison (see Table 1: Risk management)

The recommendations and requirements about the scope, design, and implementation of risk management in the four documents considered are at different levels of detail. Apart from emphasizing the importance of risk management and that consideration of trade-offs should be part of it, the HLEG does not provide detailed recommendations on which risk management procedures and processes to implement in practice. However, the European Commission is more explicit. Article 9 of the proposed AI regulation formulates requirements for a risk management system that shall be established by providers of high-risk AI systems. Like ISO, the EC sees risk management as a continuous iterative process. According to Article 9 of the proposed AI regulation, it shall comprise the identification, analysis, and evaluation (based on the continuous monitoring of the AI

14 The description in this section is based on [ISO, 2018], p. 5.

system) of risks as well as the adoption of suitable risk management measures. Moreover, the EC requires testing procedures that include tests against preliminarily defined metrics and probabilistic thresholds. Compared to this, the processes and procedures for risk management required in the AIMS Draft cover the requirements in Article 9 of the proposed AI regulation well. However, the AIMS Draft also does not provide more specific guidance than the EC on how the respective processes shall be designed or implemented. For processes regarding the identification, analysis, and evaluation of risks, the AIMS Draft refers to ISO/IEC 23894, the international standard for AI risk management [ISO/IEC, 2021b], while the issue of validation, verification, and monitoring of AI systems is treated in the controls.

In contrast, the BSI formulates more concrete requirements for several aspects of risk management, including, for instance, monitoring and assessment of possible threats and attacks with respect to integrity, availability, confidentiality and malfunction or misuse, consolidation of threat scenarios, and handling of AI specific security incidents. Still, the focus of the AIC4 catalog is on security threats, so it does not fully cover the broad spectrum of impacts of AI systems that are to be considered according to the HLEG and the EC.

When comparing the risk management requirements in the four documents, their different conceptions of risk could be problematic in the sense that, even if the procedural recommendations and requirements are the same, their implementation may yield seriously different results, depending on the conception of risk applied. To illustrate with an exaggerated example, let's assume that a provider of an AI system understands risk in terms of potential effects on its business goal, which is profit maximization. The provider could have processes in place for the identification, analysis, and evaluation of risks. Let us assume that one risk that has been identified is that users could be discriminated against by the system and that the provider would have to pay compensation as a result. The provider could set up a budget for discrimination cases, from which compensation would be paid. If this is economically more profitable for the provider than taking measures to remedy the discrimination, the provider would thus successfully manage risks with regard to achieving corporate goals. However, this example of risk management is complementary to risk management as required by the HLEG and the EC, which aim, among other things, to prevent harm to individuals and society, such as discrimination.

While the AIMS Draft does not relate its definition of risk directly to organizational goals, neither does it relate risk directly to the health, safety, and fundamental rights of individuals the way the HLEG and the EC do. The AIMS Draft addresses this discrepancy in the controls. There, it recommends fairness, security, safety, privacy, transparency and explainability, accountability, availability, maintainability, availability, quality of training data, and AI expertise as possible AI-related organizational objectives when managing risks. It thus provides a direction for risk management to work towards meeting the trustworthiness requirements of the European Commission. Moreover, in its controls, the AIMS Draft prominently addresses the notion of the impact of AI systems which, similar to risks, "can include both benefits and negative impacts or harms". With regard to impact, and apart from the impact on the organization itself, a clear recommendation is made to explicitly consider impacts to individuals and society, for instance concerning privacy, transparency, automated decision making, fairness, health, safety, culture, values, and the environment. Apart from the fact that the HLEG additionally considers impacts on fundamental rights, democracy, work, and skills, and that the EC emphasizes that specific consideration shall be given to whether the AI system is likely to be accessed by or have an impact on children, the AIMS Draft thus cuts across the majority of the aspects that THE HLEG and the EC associate with their view of risk. Further, the AIMS Draft requests organizations to integrate the process for assessing the impact of AI systems in their risk management approach. Even more, "conducting impact assessments" is recommended as a potential topic-specific policy to provide additional guidance for management.

To summarize, risk management, being an integral part of the decision-making of organizations, is a decisive factor for the trustworthiness of providers or users of AI systems. Risk management ranges from the discussion of technical dimensions (see **Section 2.3.2**), which build the technical basis for mitigation and control of AI risks, to the definition of AI policies, which should be informed by the risk environment of the organization. As illustrated in this section, different views of risk are reflected in the documents considered. Because the AIMS Draft recommends the integration of AI impact assessment into the risk management approach, it sets a suitable framework for addressing respective requirements by the upcoming AI regulation.

Table 1: Risk management

	Risk management
HLEG ALTAI	*Fundamental Rights impact assessment* *Requirement #4:* - stakeholder participation in the design and development of an AI system; stakeholders should be consulted after deployment, for instance to give feedback *Requirement #6:* - evaluation of potential negative impacts of the AI system on the environment - assessment of societal impact and impact on democracy *Requirement #7:* - ability to report on actions or decisions that contribute to the AI system's outcome - process for third parties (e.g., suppliers, end-users, subjects, distributors/vendors, or workers) to report potential vulnerabilities, risks, or biases in the AI system
European Commission Proposed regulation on AI	*Article 9:* - 'risk management system shall consist of a continuous iterative process run throughout the entire life-cycle of a high-risk AI system, requiring regular systematic updating. It shall comprise the following steps: (a) identification and analysis of the known and foreseeable risks associated with each high-risk AI system; (b) estimation and evaluation of the risks that may emerge when the high-risk AI system is used in accordance with its intended purpose and under conditions of reasonably foreseeable misuse; (c) evaluation of other possibly arising risks based on the analysis of data gathered from the post-market monitoring system referred to in Article 61; (d) adoption of suitable risk management measures in accordance with the provisions of the following paragraphs.' - risk management shall include suitable testing procedures. 'Testing shall be made against preliminarily defined metrics and probabilistic thresholds that are appropriate to the intended purpose of the high-risk AI system.' - 'specific consideration shall be given to whether the AI system is likely to be accessed by or have an impact on children'
BSI AIC4 catalog	*Security and robustness:* - continuous assessment of security threats and countermeasures (monitoring and assessment of possible threats and attacks with respect to integrity, availability, confidentiality and malfunction or misuse; consolidation in threat scenarios) - risk exposure assessment (threat models, analyze probability and impact of occurrence) - regular risk exposure assessment (regular re-evaluation of security threats, also in case there are new threats) - residual risk mitigation (in case the residual risk is still unacceptable) *Reliability:* - handling of AI specific security Incidents (document and address incidents, consolidate them into new threat scenarios) - backup and disaster recovery (policies and instructions, for back-up management)
ISO/IEC JTC 1/SC 42/WG 1 ISO/IEC WD 42001	*Controls:* - define, document, and implement responsible AI development processes that consider at what stages impact assessments should be performed - impact assessment process should integrate the concepts of impact to individuals and society as well as to the organization (impact on physical as well as intangible assets) itself - document impact assessments on the environment, health, and safety of society at large and norms, traditions, culture, and values - understand impacts of use of data (privacy, security, transparency) - impact assessment should be integrated into the risk management approach of the organization

2.3.1.2 Resources

Definition
The planning and allocation of resources (human or technical) are of practical relevance for the implementation of trustworthy AI.

Comparison (see Table 2: Resources)
The table shows that the planning and management of resources are addressed only partly by the HLEG, the EC, and the BSI. Especially, they leave it open to the developer or provider of the AI system how to plan and allocate resources in order to fulfill the other (also technical) trustworthiness requirements. In contrast, allocation and planning of resources is a prominent element of any management system. Consequently, the AIMS Draft pays explicit and dedicated attention to assure that proper handling of resources is performed in order to achieve the trustworthiness objectives.

Table 2: Resources

	Resources
HLEG ALTAI	*Requirement #3:* - Data Protection Officer (DPO) *Requirement #5 (from Ethics Guidelines):* - hiring from diverse backgrounds, cultures, and disciplines *Requirement #7:* - internal audits
European Commission Proposed regulation on AI	*Article 17:* - resource management, including security of supply related measures
BSI AIC4 catalog	*Reliability:* - resource planning for development
ISO/IEC JTC 1/SC 42/WG 1 ISO/IEC WD 42001	*Main part:* - when planning how to achieve its AI objectives, the organization shall determine what resources will be required - determine the necessary competence of persons doing work related to the AI performance - ensure that these persons are competent, or take actions to acquire necessary competence *Controls:* - 'AI Expertise' as possible organizational objective when managing risks - ensure availability of data resources, computing and system resources and human resources - define, document, and implement development processes that consider expertise required and/or training of developers

2.3.1.3 Accountability

Definition
Accountability refers to the need to define roles and responsibilities for addressing the requirements regarding the trustworthiness of AI development and the use of AI. This is particularly important for structuring an organization as well as for supporting the certification of organizations.

Comparison (See Table 3: Accountability)
While the HLEG, the proposed EU regulation on AI, and the AIC4 standard all briefly mention accountability as an important element of assuring the trustworthiness of AI, the AIMS draft goes into much more detail. This reflects the focus of the AIMS, as, similar to "resources", clear definitions of roles and responsibilities are a central element of management. It should be pointed out that the issue of "scattered responsibilities", as explained in **Section 1.2.1**, is, in particular, well recognized by the AIMS Draft allowing for tracing responsibilities across different stakeholders.

Table 3: Accountability

	Accountability
HLEG ALTAI	*Requirement #7:* - ensure responsibility for the development, deployment and/or use of AI systems - when adverse impact occurs, there should be provided for accessible mechanisms that ensure adequate redress - risk training, information about legal framework applicable to the AI system - AI ethics review board or similar mechanism *Requirement #3:* - possibility to flag issues with view to privacy and data protection *Requirement #5:* - possibility to flag issues with respect to discrimination or unfair bias
European Commission Proposed regulation on AI	*Article 17:* - an accountability framework setting out the responsibilities of the management and other staff with regard to all aspects listed in this paragraph (this refers to the quality management system) *Article 62*: - reporting of serious incidents and of malfunctioning to market surveillance authorities
BSI AIC4 catalog	*Reliability:* - resource planning for development *Security and robustness:* - implementation of countermeasures: The suitability of implemented countermeasures as well as residual risks must be formally accepted by the risk owner. In case the risk owner does not accept the remaining level of risk, SR-07 must be considered.
ISO/IEC JTC 1/SC 42/WG 1 ISO/IEC WD 42001	*Main part:* - when planning how to achieve its AI objectives, the organization shall determine who will be responsible - internal audits of the AI management system *Controls:* - 'Accountability' as potential guiding objective for responsible AI development - ensure that organization understands its responsibilities and remain accountable for these - policy for AI which should be informed by regulations, legislation, and contracts - processes for responsible use of AI (laws and regulations applicable to the organization) - ensure that responsibilities in AI lifecycle are allocated between organization, its partners, suppliers, customers and third parties - multi-stakeholder approach to development - require suppliers to implement necessary controls - asset and resource management - defined roles and responsibilities for development and performance - defined roles and responsibilities for security - defined roles and responsibilities for privacy - defined roles and responsibilities for human oversight - impact assessments and risk assessments - topic-specific policies, for instance for conducting impact assessments or for AI system development - ISO/IEC DIS 38507 describes how the governing body is accountable for the actions and decisions of the organization; those policies should be adjusted if the organization intends to use AI (wrt policy areas such as data, compliance, risk) [ISO/IEC, 2021d]

2.3.1.4 Data governance

Definition
Data governance unites the various aspects related to the activity involving data. In particular, data has to be assembled, preprocessed, and validated. All the corresponding processes should be predefined, keeping in mind the objectives of the trustworthy AI system development, and each activity should be documented.

Comparison (see Table 4: Data governance)
Data governance settlement is a requirement that is expressed in all the compared documents. In general terms, it relates to the infrastructure of the data processing for an AI system. Since any machine learning model is trained/tested using data samples, that is a mandatory part of the development process. All three documents (HLEG, EC, and BSI) emphasize the importance of having data management processes in place for use in AI systems: collection and quality check, and integrity check. The HLEG emphasizes the tests, documentation, and protocols for data usage, while EC mentions preprocessing and labeling, and problems of biases. BSI has the most thorough separation of the requirements in different stages: development, operation, data selection, annotation, and repeating quality assessment. The HLEG and the BSI recommend authorizing access to the data.

The AIMS Draft recommends taking training data expectations into account when the development process is being defined and implemented. It also includes the full description of the control for data governance inside the management system for the different stages of data usage. The availability and quality of the data are also included as a possible objective for the AIMS formulation.

Table 4: Data governance

	Data governance
HLEG ALTAI	*Requirement #3:* - oversight mechanisms for data processing - compliance with relevant standards for data management and governance - "data governance that covers the quality and integrity of data used" [HLEG, 2019] - tests and documentation of datasets for quality & integrity - data protocols to govern data access
European Commission Proposed regulation on AI	*Article 10:* - appropriate data governance and management practices (collection, preprocessing, labelling, identify shortcomings, examination of biases, …)
BSI AIC4 catalog	*Data quality:* - data quality requirements for development - data quality requirements for operation - data annotation (define requirements) - data quality assessment (regular checks of data quality) - data selection (based on defined assessment requirements, documentation of selection process) *Data management:* - data management framework - data access management (access authorization) - traceability of data sources (document the origin) - credibility of data sources (assess credibility and usability, ensure credibility e.g., by encryption)
ISO/IEC JTC 1/SC 42/WG 1 ISO/IEC WD 42001	*Controls:* - 'availability and quality of training data' as possible organizational objectives when managing risks - ensure availability of data resources - define, document, and implement development processes that consider training data expectations and rules - determine and document selection of data (define requirements for data quality, type, amount for training, source, user demographics) - ensure that data quality requirements are met (measure quality of data)

2.3.2 Technical dimensions

In this sub-section, we consider the technical dimension of the trustworthiness of AI systems, structured in line with the guidelines of the Fraunhofer IAIS audit catalog for trustworthy AI. These technical dimensions describe the areas in which AI systems encounter other challenges than classical software systems. Even if the AIMS Draft is restricted to the definition of process and strategies within an organization, it clearly refers to those technical dimensions as well, as they result in requirements on the organization. Hence it makes sense to check whether the AIMS Draft is covering the same issues as defined in the other documents. It should be noted that the EU proposed regulation on AI foresees these requirements for high-risk AI systems only.

2.3.2.1 Reliability

Definition
Reliability is a collective term for different aspects of the technical quality of an AI system. This includes the accuracy of the AI system, reliable functioning of the AI system under small disturbances of the input (robustness), and interception of errors where otherwise correct processing is not expected, in particular by suitable uncertainty evaluation of the outputs.

The risk area reliability under normal conditions is intended to ensure that the AI system operates with an appropriate degree of correctness/accuracy. This requirement is related to issues such as whether the AI system was developed and tested according to best practices and whether high-quality representative data were used in the process. Maintaining the quality of an AI system that continues to learn during operation is a particular challenge.

Comparison (see Table 5: Reliability under normal conditions)
With respect to reliability under normal conditions, there emerge three requirement areas that are treated in varying degrees of detail in the four documents considered: performance, data quality, and testing/validation methods.

A common feature of all the documents is that they demand an appropriate degree of performance, but leave the choice of evaluation metric(s) to the provider. In particular, the BSI and the ISO WG demand that, in addition to defining relevant KPIs and performance metrics, the organization itself should define target values for the metrics and release criteria that match the organization's goals.

An important factor for the quality of an AI system is the quality of the data that it is trained on and that it operates on. Accordingly, data quality is addressed in all the documents considered. While the requirements by the HLEG and the EC in this regard are rather high-level, the AIC4 catalog breaks them down further and explicitly distinguishes between data quality requirements for development, testing, and operation. The draft AI management system, on the other hand, does not indicate concrete, technical quality requirements for data, but rather treats the topic on a conceptual level. It names "availability and quality of training data" as possible organizational objectives when managing risks. Further, it demands that an organization considers requirements on training data, measures the quality of data, and ensures that the defined requirements are fulfilled.

Regarding validation and testing procedures, the EC sets requirements for post-market monitoring but does not go into detail regarding the development process (apart from training data quality). While providers of high-risk AI systems are required to prepare technical documentation that includes the development steps and the verification and validation methods used, the EC does not give a concrete indication of which measures and precautions are sufficient to demonstrate conformity during development.

The AIC4 catalog, on the other hand, addresses requirements with respect to training, validation, and testing. In the description of its criteria as well as in their supplementary information, the AIC4 catalog gives, compared to the other documents, a more concrete indication of what its requirements mean and how they can be implemented in practice. For example, it demands that under-/overfitting should be addressed during model training and that validation should be performed on a so-called "golden" dataset. Moreover, it explicitly considers automated ML frameworks. In contrast, the draft AI management system by ISO does not specify concrete technical measures regarding the development of an AI system. However, it contains controls that guide organizations in setting up responsible development processes. So, it names "reliability" as a possible guiding objective for responsible AI development. Further, it demands that organizations consider key aspects of responsible development, such as evaluation metrics and release criteria, requirements for development, verification and validation measures, and a deployment plan. The drafted AI management system demands that organizations concretize those aspects themselves, adapt them to their own goals and conditions, implement them, and ensure that all defined requirements are met.

Table 5: Reliability under normal conditions

	Reliability under normal conditions
HLEG ALTAI	*Requirement #2:* - ensure sufficient level of accuracy - reliably behave as intended - relevant, representative data of high quality - verification and validation methods to evaluate and ensure different aspects of reliability - processes for the testing and verification of the AI system's reliability
European Commission Proposed regulation on AI	*Article 15:* - achieve appropriate level of accuracy - appropriate mitigation measures for 'Feedback loops' during continuous learning *Article 10:* - training, validation and testing data sets shall be relevant, representative, free of errors, and complete *Article 17:* - techniques, procedures and systematic actions for the design, design control, design verification, development, quality control and quality assurance - examination, test, and validation procedures to be carried out before, during and after the development
BSI AIC4 catalog	*Performance and functionality:* - definition of performance requirements - model selection and suitability - model training and validation - business testing - additional considerations when using automated Machine Learning *Data quality:* - data quality assessment (regular checks of data quality) - preparation of training, validation and test data
ISO/IEC JTC 1/SC 42/WG 1 ISO/IEC WD 42001	*Controls:* - 'Reliability' as potential guiding objective for responsible AI development - impact assessment should consider predictable failures and their potential impact - specify the business need - document requirements for development and data pre-processing needs - understand the impact of use of data (representativity of training data) - define, document, and implement responsible AI development processes that consider testing requirements, training data expectations and rules, and release criteria - define and document verification and validation measures (test methodologies, test data, release criteria, relevant KPIs and evaluation metrics) - design and develop the AI system according to the organizational objectives - deployment plan - ensure that release criteria are met before deployment

Definition (continued)
The **robustness** of an AI system refers to the quality with which the AI component (especially the ML model) processes disturbed input data that would be processed error-free under normal circumstances. Examples of such deviations are image distortions, noise from sensors, or general inaccuracies in the data. A special class of disturbances is the so-called "adversarial examples", where small deviations of the input data are generated in such a way that the output seriously deviates from the expected result. These examples can be the consequences of targeted attacks (which are treated in the security & safety dimension), but they are generally an expression of model weaknesses.

However, robustness, which aims at the correct functioning of the AI component, can usually not be achieved for an arbitrarily large input space. If failure of the AI component is foreseeable or unavoidable, or if a greater than usual number of incorrect outputs would lead to unacceptable risk, errors must be intercepted to limit potential damage. A key measure for this is the reliable detection of problematic input data for which meaningful processing cannot be expected or whose unchecked processing may lead to unjustifiable risk. The risk area **interception of errors** in the reliability dimension is closely related to the risk area functional safety since functional monitoring plays an essential role in both risk areas. From a technical point of view, they differ in that "interception of errors" refers to (AI-specific) detection mechanisms at the level of the AI component, whereas functional safety addresses classical mitigation measures. Especially, a detection mechanism at the model level can also trigger follow-up reactions from the area of functional safety, such as a roll-back to a previous version of the model.

Uncertainty about the correctness of an output is an intrinsic property of data-driven systems. ML-based AI systems often do not provide an unambiguous answer, but they can be viewed as a probabilistic function. Uncertainty evaluation is often an important component for safety reasoning. Thus, an uncertain result should be confirmed by another component of the overall system or a human.

Comparison (see Table 6: Reliability: Robustness, error handling, uncertainty)
When considering reliability in terms of how an AI system performs under challenging conditions and how it handles errors, it should be taken into account that a uniform terminology has not yet emerged and that terms like 'reliability', 'robustness' or 'resilience' are differently connotated in the documents examined. Table 6 compares requirements regarding 'robustness', 'interception of errors' and 'uncertainty' as defined in the introduction of this section. Moreover, the distinction as to whether errors are avoided/handled by the AI component or by its embedding is not made explicit in any of the four documents, so that some requirements can be interpreted as belonging to both the reliability dimension and the safety dimension.

Although a concrete, quantitative description of this requirement hardly seems possible, all four documents emphasize robustness as an important aspect of trustworthy AI.

The HLEG and EC require a consistent performance of the AI system on a range of inputs that should be tested accordingly. Moreover, the EC explicitly mentions "adversarial examples" and model flaws to be considered by technical solutions.

On the other hand, the AIC4 catalog sees robustness as strongly related to security. Its testing requirements focus on attacks that are based on the violation of data integrity and the exploitation of model vulnerabilities. Here, the AIC4 is again more concrete than the other documents about potential attack scenarios and countermeasures, among which it also includes privacy and cybersecurity methods. With regard to the interception of errors, the AIC4 is also the only one among the four documents which explicitly demands checks of user requests to detect malicious inputs.

The drafted AI management system suggests robustness, in terms of consistent performance on typical input data, as a possible organizational objective when managing risks. However, in contrast to the other documents, the drafted AI management system does not provide for any concrete technical control or requirement related to the implementation of robustness apart from testing. In particular, it does not require organizations to have processes in place for the detection of malicious inputs, as, for example, demanded in the AIC4.

Both the ALTAI and the AIC4 catalog address the uncertainty of AI systems. They require an indication as to how likely errors of the AI system are. In particular, the AIC4 demands a sensitivity analysis of the performance metric against subparts of the input space. Neither the proposed regulation nor the drafted AI management system address this issue. Especially, the drafted AI management system does not suggest organizational procedures for handling low confidence results, as demanded by the HLEG.

Table 6: Reliability: Robustness, error handling, uncertainty

	Robustness	Interception of errors	Uncertainty
HLEG ALTAI	*Requirement #2:* - robustness when facing changes - work properly with a range of inputs and in a range of situations[15] - verification and validation methods to evaluate and ensure different aspects of reliability		*Requirement #2:* - system should indicate how likely errors are[16] *Requirement #4:* - procedures for handling low confidence of results
European Commission Proposed regulation on AI	*Article 15:* - consistent performance throughout the lifecycle - technical solutions to address AI specific vulnerabilities - measures to prevent and control for 'adversarial examples' or model flaws	*Article 15:* - technical solutions to address AI specific vulnerabilities - resilience against errors, faults or inconsistencies that may occur within the system or its environment	
BSI AIC4 catalog	*Security and robustness:* - testing of learning pipeline robustness (tests to measure the risk associated with data integrity, simulate attacks based on manipulated data) - testing of model robustness (tests to exploit model vulnerabilities) - implementation of countermeasures	*Security and robustness:* - continuous assessment of security threats and countermeasures (monitoring and assessment of possible threats and attacks with respect to integrity, availability, confidentiality and malfunction or misuse, consolidate in threat scenarios) - risk exposure assessment (threat models, analyze probability and impact of occurrence) - regular risk exposure assessment (regular re-evaluation of SR-02, or in case there are new threats) - residual risk mitigation (in case the residual risk is still unacceptable) - implementation of countermeasures (e.g., anomaly detection) *Reliability:* - monitoring of model requests (to detect malicious requests)	*Performance and functionality:* - definition of performance requirements (including confidence levels)

15 From [HLEG, 2019b] key requirement #2, section "reliability and reproducibility".

16 From [HLEG, 2019b] kex requirement #2, section "accuracy".

	Robustness	Interception of errors	Uncertainty
ISO/IEC JTC 1/SC 42/WG 1 ISO/IEC WD 42001	*Controls:* - 'Robustness' as possible organizational objective when managing risks - performance assessment methodologies which may require controlled introduction of erroneous or spurious data	*Controls:* - processes for response to errors and failures of the AI system - processes for repair of errors and failures - processes for updating the system	

2.3.2.2 Safety and security

Definition
Mitigating risks in terms of safety and security is a major challenge that must be solved by a combination of classical measures (from functional and cyber security) and AI-specific methods - and thus in close coordination with the technical measures for reliability.

Functional safety refers to the goal of ensuring that an AI system does not lead to conditions where human life, health, property, or the environment are at risk in case of malfunction.

The goal of **security**, on the other hand, is to prevent undesired external influences on the system, for example by humans or other machines. New attack vectors have evolved, particularly on the integrity and availability of AI applications, such as the spying out of the model or sensitive training data, or the targeted manipulation of training data, which can result in a functional change of the system. Precautions must be taken to prevent this.

Comparison (see Table 7: Safety and security)
In terms of safety, the HLEG, the EC, and the BSI consistently require resilience against errors and faults. While the BSI specifies requirements that take off on backup management, the EC and the HLEG additionally mention technical solutions, which, for example, may involve redundancy.

The draft AI management system highlights safety as a possible organizational objective for responsible development as well as for risk management. In this regard, it demands processes for response and repair of errors and failures. Although it does not go into (technical) detail, the draft does set a framework into which procedures for backup management and failsafe, as demanded in the proposed regulation, can be placed.

With respect to security, the HLEG and the EC address cybersecurity on the one hand, and, on the other, also demand that resilience against (new) AI-specific attacks and vulnerabilities needs to be achieved.

Similarly, the BSI requires appropriate handling of AI-specific security issues, whereby it explicitly includes classical security measures for the protection of data integrity, such as access management and encryption.

The ISO highlights security and availability as possible guiding objectives for development and risk management. Instead of directly referring to cybersecurity standards, as the HLEG does, the AI management system aims to achieve compliance with the organization's security policy, whereby it demands that AI-specific threats also need to be controlled. However, the draft management system does not prescribe concrete technical measures.

Table 7: Safety and security

	Functional safety	Integrity and availability
HLEG ALTAI	*Requirement #2:* - process to continuously measure and assess risks or threats such as design or technical faults, defects, outages, misuse, inappropriate or malicious use - fault tolerance should be ensured via, e.g., redundancy - failsafe fallback plans	*Requirement #2:* - compliance with relevant cybersecurity standards - assessment of potential forms of attacks - resilience against AI-specific attacks and vulnerabilities (data poisoning, model inversion, model evasion) - red-team/pentest
European Commission Proposed regulation on AI	*Article 15:* - resilience against errors, faults or inconsistencies that may occur within the system or its environment - technical redundancy solutions, which may include backup or fail-safe plans	*Article 15:* - measures to prevent and control for attacks trying to manipulate the training dataset ('Data poisoning') - resilience against attempts to alter use or performance by exploiting the system vulnerabilities - technical solutions aimed at ensuring the cybersecurity
BSI AIC4 catalog	*Reliability:* - backup and disaster recovery: - policies and instructions with safeguards to avoid loss of data and model(s) - procedures for back-up management - at least annual tests of recovery procedures	*Security and robustness:* - continuous assessment of security threats and countermeasures (monitoring and assessment of possible threats and attacks wrt integrity, availability, confidentiality and malfunction or misuse, consolidate in threat scenarios) - risk exposure assessment (threat models, analyze probability and impact of occurrence) - regular risk exposure assessment (regular re-evaluation, or in case there are new threats) - residual risk mitigation (in case the residual risk is still unacceptable) - implementation of countermeasures *Reliability:* - monitoring of model requests - handling of AI specific security incidents *Data management:* - data access management - credibility of data sources (ensure credibility e.g., encryption)
ISO/IEC JTC 1/SC 42/WG 1 ISO/IEC WD 42001	*Controls:* - 'Safety' as potential guiding objective for responsible AI development - 'Safety' as possible organizational objective when managing risks - understand the impact of use of data (security and safety threats) - processes for response to errors and failures of the AI system - processes for repair of errors and failures - processes for updating the system	*Controls:* - 'Privacy and security' as potential guiding objectives for responsible AI development - 'Security' and 'Availability' as possible organizational objectives when managing risks - understand the impact of use of data (security and safety threats) - AI systems should be compliant with the organization's security policy - ensure that AI-specific threats are addressed by existing security measures - consider security threats during the full lifecycle

2.3.2.3 Data protection

Definition
AI systems often rely on processing huge amounts of data that contain sensitive information. These can be personal data, which are protected by the General Data Protection Regulation, as well as other information worthy of protection, such as the model itself or trade secrets.

The data protection requirements for AI systems are often much higher than for conventional IT systems, as AI systems frequently combine data that was previously not linked, and AI-based processes create new opportunities for linking. Both classic cybersecurity methods such as encryption and AI-specific methods such as federated learning can be used to protect personal as well as business-related data.

Comparison (see Table 8: Data protection)
The recommendations to the data protection can be split into two areas – related to the personal data of people, that is used for training models, and related to the business data, which also includes information about the model itself. HLEG recommends estimating the impact that can be induced by an AI system on the private data of the users. All three documents suggest considering cybersecurity measures to protect the business-related private information. While HLEG does not provide details on the possibilities for implementing the data protection, the EC document suggests pseudo-anonymization and encryption. BSI concentrates on recommending the implementation of countermeasures against both types of privacy corruption. The AIMS standard draft pays attention to the privacy aspects as the goal when an AI system is developed, as well as when the properties of the data used are analyzed. In particular, privacy is the recommendation for analysis on the impacts of an AI system as well as the aspect that determines the responsibilities of the stakeholders in the case when private data is involved in the process. Nevertheless, the security aspects of the system, with relation to the privacy of the business data, are addressed rather briefly among the possible objectives of the responsible development of an AI system.

Similarly, the AIMS draft directs an organization to assure cybersecurity standards and check the privacy-related issues that appear when an AI system is being developed and delivered.

Table 8: Data protection

	Protection of personal data	Protection of business-related data
HLEG ALTAI	*Requirement #3:* - consider the impact of the AI system on the right to privacy, the right to physical, mental and/or moral integrity and the right to data protection - technical measures for data protection to achieve compliance with GDPR ('Privacy-by-design')	*Requirement #2:* - compliance with relevant cybersecurity standards - resilience against model inversion (i.e., leakage of model parameters)
European Commission Proposed regulation on AI	*Article 10:* - security or privacy-preserving measures (e.g., pseudonymization, encryption) in case special categories of personal data are processed	*Article 15:* - technical solutions aimed at ensuring the cybersecurity
BSI AIC4 catalog	*Security and robustness:* - implementation of countermeasures (for privacy, in order to prevent attacks – which corresponds to their understanding of robustness)	*Security and robustness:* - implementation of countermeasures (countermeasures against threat models derived from threat scenarios identified in SR-01: these include threats like leakage of data or model, model stealing attacks, membership inference attacks)
ISO/IEC JTC 1/SC 42/WG 1 ISO/IEC WD 42001	*Controls:* - 'Privacy' as possible organizational objective when managing risks - determine privacy impacts of the AI system - identify and comply with the applicable obligations related to PII processing - reference to controls such as those described in ISO/IEC 27701 [15]	*Controls:* - 'Privacy and security' as potential guiding objectives for responsible AI development

2.3.2.4 Transparency

Definition
The transparency of AI applications comprises various aspects, such as the technical explainability of outputs, their reproducibility, and the traceability of the AI system.

The challenge of explainability is that many AI technologies (e.g., those realized by using large neural networks) are based on mathematical models in which many thousands or millions of parameters are fitted to so-called training data. The meaning of these parameters is often difficult for humans to understand, which is why the systems are referred to as "black boxes". However, it may be essential for the responsible use of an AI application that the factors that led to its results can be understood by a human (**explainability for users**). Transparency should also be established (at a deeper technical level) **for experts** so that they can validate the system, trace possible weaknesses and causes of errors, and, if necessary, make adjustments to the system. It is also important for human oversight – a human can take over control in the case of erroneous behavior. In the case of black-box models, explainability can be achieved through downstream technical tools. In addition, several AI models are a priori interpretable by humans, such as rule-based models. Which procedure is used to realize a particular AI system depends on the individual case. In particular, it should be noted that higher comprehensibility and interpretability can be at the expense of the robustness and performance of the AI system so that careful consideration is necessary here.

A set of requirements on trustworthy AI systems deals with technical measures that assure that the decisions of the AI system can be traced. The final goal is often that the system developer or provider can approve the proper functioning of the system in a potential system audit. As the notion of auditability in the context of organizations is often referred to as the transparency of an organization and its processes for external audit, we use here the term traceability of an AI application to describe its readiness for possible (internal or external) system audits. This comprises, in particular, detailed technical documentation of the structure, development, and functioning of the AI application, as well as the data used for this purpose. These aspects can also enable operators to trace specific outputs of the AI application or the cause of errors in liability issues. The reproducibility of outputs and of the ML model itself also plays a role here.

Comparison (see Table 9: Transparency)
Explainability (transparency) for users is addressed by the HLEG, the EC, and the BSI. The recommendation is to integrate technical mechanisms in order to obtain explanations for particular decisions of the system. Moreover, the explanations should correspond to the person obtaining it: if it is a layperson, domain specialist, or developer. In particular, the EC addresses the need for the explainability for the persons who perform a human oversight function so that they can take over control in case of failures. While the HLEG emphasizes the need for explainability, especially in the case when humans' lives are affected, the BSI recommends assessing the degree of interpretability required. The BSI also recommends the processes for testing and evaluating interpretability. All three documents recommend proper documentation of the system and logging its performance, thus addressing the traceability aspect of transparency. Moreover, for the EC it is also important regarding the post-market monitoring, which is addressed in **Chapter 2.3.1**, and all the organizations show the importance of the data sources and characteristics being documented (which is also related to data governance addressed in **Chapter 2.3.1**).

The AIMS standard draft addresses the need for explainability by recommending that it is one of the goals of the AI system development and part of the impact assessment. Possibly the aspects related to the human oversight and cases of human life affected by an AI system can be addressed more directly. With respect to the aspect of the documentation for traceability, the AIMS standard includes most of the needed sides: documentation of the data, resources, and all the development stages. Record keeping and logging, however, are not mentioned explicitly, thus addressing only partly the need for transparency of an AI system for later inspections.

Table 9: Transparency

	Explainability for users	Transparency for experts	Traceability
HLEG ALTAI	*Requirement #4:* - technical processes and reasoning behind an AI system's decisions should be explained to users and affected persons to the degree possible	*Requirement #4*[17]*:* - technical explainability - suitable explanation of the decision-making process whenever an AI system has a significant impact on people's lives - explanation should be adapted to expertise of the stakeholder (e.g., regulator, researcher)	*Requirement #4:* - mechanisms and procedures for record-keeping to allow for traceability of outputs *Requirement #7:* - facilitate internal or external audits as appropriate - documentation of processes and record-keeping *Requirement #2:* - relevant data should be documented and, if appropriate, specific contexts/scenarios should be taken into account to ensure reproducibility
European Commission Proposed regulation on AI	*Article 13:* - technical measures to facilitate the interpretation of the outputs by the users	*Article 14:* - tools and methods to facilitate the interpretation of the outputs by the individuals to whom human oversight is assigned	*Article 12:* - automatic recording of events - logging *Article 11:* - a technical documentation of the AI system shall be drawn up in such a way to demonstrate compliance with the criteria in Chapter 2 of the proposed European regulation[18] *Article 61:* - post-market monitoring system shall allow the provider to continuously evaluate compliance with the requirements
BSI AIC4 catalog	*Explainability:* - assessment of the required degree of explainability - provide explanations about why a specific output was produced, as appropriate - explanations must be tailored for the recipients (such as subject matter experts, developers, users)	*Explainability:* - assessment of the required degree of explainability - testing the explainability of the service - provide explanations about why a specific output was produced, as appropriate - explanations must be tailored for the recipients (such as subject matter experts, developers, users)	*Performance and functionality:* - regular service review (logging user feedback, failures, …) *Reliability:* - logging of model requests (for backtracking failures and incidents) - backup and disaster recovery (back-ups for data and model) *Data management:* - traceability of data sources (document the origin)

17 The content in this cell is taken from [HLEG, 2019b], section on the key requirement "transparency." The ALTAI does not explicitly take up explainability for experts.

18 The criteria in Chapter 2 of the Proposal for Regulation are risk management, data and data governance, record-keeping, transparency, human oversight, accuracy, robustness, and cybersecurity. The minimal elements to be covered by the technical documentation are described in Annex IV of the proposed regulation [EC, 2021].

	Explainability for users	Transparency for experts	Traceability
ISO/IEC JTC 1/SC 42/WG 1 ISO/IEC WD 42001	*Controls:* - 'Transparency' as potential guiding objective for responsible AI development - 'Transparency and explainability' as possible organizational objective when managing risks - understand the impact of use of data (transparency and explainability aspects)	*Controls:* - 'Transparency' as potential guiding objective for responsible AI development - 'Transparency and explainability' as possible organizational objective when managing risks - understand the impact of use of data (transparency and explainability aspects)	*Controls:* - document information about the data utilized (origin, categories) - document information about types of computing resources utilized (algorithm types, evaluation methods, machine learning approaches, etc.) - document information about system resources utilized - documentation of AI system design - system architecture diagram - documentation of development process - documentation of verification and validation measures - documentation of deployment plan - document pre-processing of data

2.3.2.5 Human agency and oversight

Definition
This dimension of trustworthiness deals with the tension between human autonomy on the one hand, and the autonomy of the AI-based system on the other. The meaning of autonomy in this context is the degree of automation of the AI system; in particular, the extent to which it determines the means for achieving its goals itself.

An essential means of preserving human autonomy in the use of AI systems is human oversight. Depending on the application context and the criticality of the AI system, humans must be granted appropriate intervention and revision options.

In order for users to be able to decide and act autonomously, they must also be provided with all the necessary information. Users and those affected by AI systems should not only know that they are interacting with one (i.e., the AI system is labeled as such) but should also be adequately familiarized with the use of the system. Finally, users should be informed about the capabilities and limitations of the AI system and the risks associated with its use.

Comparison (see Table 10: Human agency and oversight)
The views on the human oversight aspect are rather different for the three organizations. The HLEG addresses the high level of human oversight requirements: a "STOP" button and the possibility of performing oversight with specifically trained people. The EC gives many more details, such as the need for a suitable interface for oversight and the need for different ways of interruption depending on the situation. The BSI puts the main emphasis on the performance tracking and corrective measures needed in case of low performance. At the same time, in relation to the information provided to the human stakeholders, all the recommendations address the provision of full documentation to users, including explanations of the automation, possible risks, purposes, etc. The BSI talks about more technical aspects, such as information about logging, training, and interpretability approaches. The HLEG and EC are more concentrated on the high-level impacts, interaction, and human rights.

The AIMS standard draft recommends having a reporting system that helps to perform oversight; the recommendations also include resource specification for human oversight as well as including it in the development phase. More technical details largely overlap with the recommendations for the audibility of an AI system. It recommends that users are provided with information about updates; the communication about the AI management system is also mentioned. Possibly, the stakeholder-oriented information dissemination might be a valuable addition.

Table 10: Human agency and oversight

	Human oversight	Information and capabilities of users
HLEG ALTAI	*Requirement #1:* - human oversight as appropriate - 'Stop button' - detection mechanisms - training for humans which oversee the system *Requirement #3:* - oversight mechanisms for data processing	*Requirement #1:* - disclose the use of AI - avoid/prevent manipulation of humans, over-reliance, disproportionate attachment, or addiction *Requirement #4:* - information about purpose, capabilities, and limitations of the AI system as appropriate - material on how to adequately use the system *Requirement #2:* - inform users about risks related to general safety - information about level of accuracy - inform end-users about the given security coverage and updates *Requirement #6:* - inform impacted workers - ensure that workers understand how the AI system operates - provide training opportunities and materials
European Commission Proposed regulation on AI	*Article 14:* - appropriate human-machine interface tools such that the AI system can be effectively overseen by natural persons - human oversight measures to detect and address dysfunction and unexpected performance - ability to decide not to use the system or otherwise override or reverse the output - ability to intervene or interrupt the operation - individuals assigned for oversight should fully understand the capacities and limitations of the system - be aware of automation bias	*Article 10:* - residual risks shall be communicated to the user *Article 13:* - instructions for use, provide information to users about: characteristics, capabilities, and limitations of performance (purpose, level of accuracy, circumstances/misuse which may lead to risks) changes to the systemhuman oversight measurestechnical measures for explicabilityexpected lifetime, maintenance, and care measures *Article 15:* - levels of accuracy and accuracy metrics in the instructions of use *Article 52:* - inform natural persons if they interact with an AI system, unless this is obvious from the context - inform natural persons if exposed to emotion recognition or biometric categorization - disclosure of ‚deep fakes' (unless authorized by law)

	Human oversight	Information and capabilities of users
BSI AIC4 catalog	*Performance and functionality:* - monitoring of performance (provider assigns personnel to continuously compute and monitor the performance metric(s) defined in PF-01. In scheduled intervals (at least quarterly) reports on the performance) - impact of automated decision-making (procedures/measures for users to update or modify automated decisions) *Reliability:* - corrective measures to the output (by authorized subjects)	*System description:* - information about goals/purpose, design, and application of the AI system - assumptions and limitations of the model - information about what users are required to do in case of security incidents (handling of AI specific security incidents) - inform about impacts of the AI system - training procedure and selection of training data - inform about logging (RE-02) - quantification and limits of the robustness of the AI system - technical limitations of implemented transparency methods (EX-01) - BI-01, BI-02 & BI-03: inform user about identified biases and possible implications and with which metric the bias was measured *Performance and functionality:* - fulfillment of contractual agreement of performance requirements (make deviations transparent to users) *Data quality:* - data quality requirements for operation (make them transparent)
ISO/IEC JTC 1/SC 42/WG 1 ISO/IEC WD 42001	*Controls:* - resources necessary can include roles related to human oversight - define, document, and implement development processes that consider human oversight requirements - document potential effects of automated decision making	*Main part:* - make sure that workers are competent (e.g., by education, training, or experience) - determine internal and external communication relevant to the AI management system *Controls:* - support process for users during operation - information to users about updates

2.3.2.6 Fairness

Definition

Fairness in the sense of avoiding unjustified discrimination by AI systems is an important requirement for trustworthiness. "Unfair" behavior of AI-based decision support systems may result from the fact that they often learn their behavior from historical training data. Such training data may be unbalanced or biased. For example, if the data contains a bias against a particular group, an AI-based decision support system may adopt it. Another source of discrimination by AI systems is the statistical underrepresentation of certain groups of people in the training data, which results in lower performance of the AI system on that subset of the dataset and thus may also lead to unfair behavior. Accessibility refers to the requirement that specific groups, such as persons with disabilities, are not discriminated against.

To operationalize fairness, the first step is to find a quantifiable concept of fairness. Furthermore, technical countermeasures must be taken to ensure that bias is not continued or even reinforced in the model.

AI systems should be user-centric and designed in a way to be accessible to all people, independent of their age, abilities, gender, or characteristics. This makes an AI system fair in a way that it can be used.

Comparison (see Table 11: Fairness)
Unfairness in an AI system is twofold: unfairness in the treatment of the objects of decision and unfairness in the ability of stakeholders to use the AI system (accessibility). While the first aspect is addressed by all the three documents, the second aspect is mentioned only in the HLEG recommendations. While the HLEG recommends that the fairness characteristic is selected and assessed directly, the EC and BSI mainly recommend checking the data and model for the presence of biases, without addressing the identification of unfairly treated persons. With regard to accessibility, the HLEG and EC recommend checking whether it is possible for end-users with restricted abilities to use the system.

As with previous technical requirements, fairness is one of the recommended goals of a trustworthy AI development, according to the AIMS standard draft. Also, the processes related to the data quality checks include recommendations for addressing bias checks, as well as considering bias in the impact assessment checks. Accessibility is included as an objective in both the responsible development and use of an AI system.

Table 11: Fairness

	Non-discrimination	**Accessibility**
HLEG ALTAI	*Requirement #5:* - unfair bias in the AI system should be avoided - appropriate definition of fairness - identification of subjects that could be (in)directly affected by the AI system - diverse and representative data - adequate measures in the algorithm design against potential biases - measuring and testing of the applied definition of fairness	*Requirement #5:* - accessibility to a wide range of users - universal design principles - particular consideration or involvement of potential end-users with special needs - assessment of risks of unfairness onto user communities
European Commission Proposed regulation on AI	*Article 10:* - examination of datasets in view of possible biases *Article 15:* - possibly biased outputs shall be addressed with appropriate mitigation measures	*Article 69:* - codes of conduct may further contain a voluntary commitment to meet additional requirements, provided that the codes of conduct set out clear objectives and contain key performance indicators to measure the achievement of those objectives. Such additional requirements may relate to environmental sustainability, accessibility to persons with disability, stakeholders participation in the design and development of the AI systems, diversity of the development teams.
BSI AIC4 catalog	*Performance and functionality:* - model training and validation (absence of bias, trade-off between bias mitigation and performance) *Bias:* - assessing the level of bias (evaluation of data and model with fairness metrics) - mitigation of detected bias	

	Non-discrimination	Accessibility
ISO/IEC JTC 1/SC 42/WG 1 ISO/IEC WD 42001	*Controls:* - 'Fairness' as potential guiding objective for responsible AI development - 'Fairness' as possible organizational objective when managing risks - impact assessment should consider bias and relevant demographic groups - performance criteria should consider all aspects of operational performance that could impact stakeholders - ensure that data does not have statistical biases and does not lead to biased output	*Controls:* - 'Accessibility' as potential guiding objective for responsible AI development - 'Accessibility' as potential guiding objective for responsible AI use - impact assessment should consider quality of service impacts (such as based on demographics)

2.3.2.7 Handling of trade-offs

Definition

It should be noted that the concrete requirements to be placed on an AI system in order to fulfill trustworthiness depend to a high degree on the technology used and the context of the use of the AI system. One challenge here is that the different fields of action of trustworthiness cannot be assessed independently of each other, but rather conflicting goals may arise. For example, an increase in performance, such as recognition performance in object recognition on image data by so-called deep neural networks, can be at the expense of traceability, or an increase in transparency can lead to new attack vectors in terms of IT security.

Comparison (see Table 12: Handling of trade-offs)

The main goal of the recommendations with respect to trade-offs between different technical requirements is to make clear their presence and analyze possibilities and the consequences of the choice made. So, the HLEG and EC discuss the need to document and examine any trade-offs in an AI system. The BSI has a more limited and concrete recommendation to address trade-offs between performance and fairness, as well as performance and transparency (explainability).

The AIMS standard draft does not address the handling processes and related documentation for trade-offs. We would recommend mentioning the need for a discussion of trade-offs in the AIMS as well.

Table 12: Handling of trade-offs

	Discussion of trade-offs
HLEG ALTAI	*Requirement #7:* - trade-offs should be explicitly acknowledged and evaluated in terms of their risk to safety and ethical principles, including fundamental rights. Any decision about which trade-off to make should be well reasoned and properly documented
European Commission Proposed regulation on AI	*Annex IV, 2b):* - decisions about any possible trade-off made regarding the technical solutions adopted to comply with the requirements [in Chapter 2 of the document] should be described
BSI AIC4 catalog	*Performance and functionality:* - trade-offs between performance and bias mitigation *Explainability:* - trade-off between transparency and performance
ISO/IEC JTC 1/SC 42/WG 1 ISO/IEC WD 42001	

2.3.2.8 Post-market monitoring

Definition
Another difference between ML-based models and classic software is that the operational environment and the AI systems can in principle evolve and change during operation. A distinction is made here between so-called "concept drift" and "model drift". In the case of "concept drift", the distribution of the data normally to be expected and/or the real-world context represented by the model change in the operational environment, so that the model reflects this correspondingly less correctly and must be retrained. Thus, an AI-based pedestrian-recognition system that has been trained in its learning process only with images of pedestrians without mouth-to-nose coverings will struggle to achieve the same performance for images of pedestrians wearing such mouth-to-nose coverings. "Model drift", on the other hand, refers to the possibility that the training process of the ML model continues in active operation, for example, by the system continuing to learn through user feedback.

Moreover, multiple aspects of the system performance can be assessed only through the life performance, especially in the cases with expensive/not available data. Thus, special handling of the post-production should be in place for an AI system to be trustworthy, providing support to the end users in cases of failures.

It is worthful noting that post-market monitoring requires organizational as well as technical measures to be taken.

Regarding organizations, processes responsibilities must be defined and documented, and resources must be allocated that perform the post-market monitoring. Technically, measures must be foreseen within the AI system that produce metrics during the operation of the AI system which allow for the assessment of all the technical dimensions discussed above. Obviously, these two aspects of post-market monitoring are mutually dependent: Establishing a monitoring process to monitor a system that does not produce any metrics would be as useless as a system producing relevant metrics that are not used by anyone.

Comparison (see Table 13: Post-market monitoring)
The HLEG addresses the requirements of the system with respect to the model and data drift, and also recommends monitoring the performance and adherence of the system. The EC addresses the issue of post-market monitoring from a higher level and recommends collecting and analyzing data related to the system performance, which will allow for continuous checks against the requirements. The BSI addresses all the questions related to post-production security and safety but also recommends logging for improving performance and reliability, as well as checking for biases in data.

The AIMS standard draft emphasizes the need for post-market monitoring for detecting errors and possible ways to improve performance (the need to retrain). Possibly, recommendations for continuous security and safety checks, as well as data quality checks, might be added for completeness.

Table 13: Post-market monitoring

	Post-market monitoring
HLEG ALTAI	*Requirement #2:* - monitoring of the level of accuracy - continuous evaluation of reliability - processes to continuously measure and assess risks related to general safety - risks of continual learning should be addressed *Requirement #4:* - measures to continuously assess the quality of the input data and the output *Requirement #7:* - process to discuss and continuously monitor and assess the AI system's adherence to this Assessment List for Trustworthy AI
European Commission Proposed regulation on AI	*Article 61:* - post-market monitoring system as proportionate, based on a post-market monitoring plan - shall actively and systematically collect, document and analyze relevant data on the AI system's performance - shall allow the provider to continuously evaluate compliance with the requirements

	Post-market monitoring
BSI AIC4 catalog	*Security and robustness:* - continuous assessment of security threats and countermeasures (monitoring and assessment of possible threats and attacks wrt integrity, availability, confidentiality and malfunction or misuse, consolidate in threat scenarios) - risk exposure assessment (threat models, analyze probability and impact of occurrence) - regular risk exposure assessment (regular re-evaluation of SR-02, or in case there are new threats) - residual risk mitigation (in case the residual risk is still unacceptable) *Reliability:* - handling of AI specific security incidents (consolidate security incidents into new threat scenarios) *Performance and functionality:* - monitoring to performance (provider assigns personnel to continuously compute and monitor defined performance metric(s), reports on the performance in scheduled intervals – at least quarterly) - fulfillment of contractual agreement of performance requirements (request re-training if necessary, make deviations transparent to users) - business testing (tests before deployment + on regular basis with golden dataset) - continuous improvement of model performance (if necessary, retraining after PF-02) - regular service review (regular review of user feedback and failure reports by subject matter expert, action plan for necessary changes) *Reliability:* - logging of model requests (for backtracking failures and incidents) - monitoring of model requests (to detect malicious requests, on a regular basis) *Data quality:* - data quality assessment (continuously) *Bias:* - assessing the level of bias (evaluation of data and model with fairness metrics) - mitigation of detected bias - continuous bias assessment
ISO/IEC JTC 1/SC 42/WG 1 ISO/IEC WD 42001	*Main part:* - determine what needs to be monitored and measured and how - evaluate the AI performance *Controls:* - 'Maintainability' as possible organizational objectives when managing risks - define elements for operation and monitoring - performance monitoring against technical performance criteria (performance is defined very general; it can also mean fairness for example) - monitoring for errors and failures - identify need for re-training (monitoring concept or data drift) - update processes, processes for reporting and repair of issues

3. Development of a Certification Approach

An essential step in building trust in AI systems is to demonstrate their quality in a credible and substantiated manner. A proven approach to do so in domains other than AI is conformity assessments against broadly recognized standards. The previous chapters have shown that requirements for trustworthy AI have also been prominently discussed and corresponding standardization and regulatory initiatives have become active. It was further elaborated that trustworthiness in the case of AI cannot exclusively refer to technical properties, but also essentially depends on how well organizations are positioned to solve challenges in the face of AI and to manage specific risks. Having discussed trustworthiness requirements and management systems as a possible framework to implement these, this chapter now looks at the issue of attesting to conformity. In particular, the concept of certification is addressed as a building block for trustworthy AI. Here, in accordance with the previous explanations, certification is considered both with regard to management systems and AI systems themselves.

Certification denotes the final stage of conformity assessment, where - based on neutral inspection or testing, and in case of positive results - the conformity of an object with previously defined requirements is attested. The precise ISO definitions of certification and related concepts are given in **Section 3.1**. Several entities can, in principle, be the object of conformity assessment and certification in particular, for example, products, processes, services, systems, persons or organizations, and any combination thereof. Normally, certification involves the issuing of a written statement (certificate) that assures that specified requirements are met. A key motivation behind certification is that a certificate, being issued by an independent and accredited body, gives third parties a well-recognized indication of the quality or desired features of the certified object. This builds trust, contributes to branding, and thus creates competitive advantages. As illustrated by the example of management systems in **Section 2.1**, demonstrating conformity with established standards can also ease compliance dependencies and can even be a regulatory requirement in itself.

With the recently published proposal for AI regulation by the European Commission, it is becoming apparent that conformity assessments for AI will soon be established as part of mandatory approval and supervisory procedures for the European market. This concerns "high-risk" AI systems which, according to the EC's definition, include a large number of applications that are already an integral part of our everyday lives. On the one hand, the proposed regulation demands that providers demonstrate the conformity of such AI systems with (technical) requirements. On the other hand, providers of high-risk AI applications shall have a quality and risk management system in place. Depending on the type of AI system, conformity can either be declared by the provider itself or, for particularly critical applications, a notified body will have to be involved to certify conformity. However, certification schemes are needed to put certification into practice, especially when it comes to certifying specific properties of AI systems. **Section 3.2** considers approaches to certification of both AI systems and AI management systems. AIMS certification would add value to organizations that want to generate evidence that they are well positioned to solve challenges in the face of AI and to manage specific risks.

3.1 General terms

The ISO/IEC standard for conformity assessment (ISO/IEC 17000:2020 [ISO/IEC, 2020b]) defines several concepts related to conformity assessment, such as inspection, testing, and certification. The following terms and definitions are taken from this standard. We will make use of these terms in **Section 3.2**.

"conformity assessment: demonstration that specified requirements are fulfilled

- Note 1 to entry: The process of conformity assessment […] can have a negative outcome, i.e., demonstrating that the specified requirements are not fulfilled.
- Note 2 to entry: Conformity assessment includes activities […] such as but not limited to testing, inspection, validation, verification, certification, and accreditation.
- Note 3 to entry: Conformity assessment is […] a series of functions. Activities contributing to any of these functions can be described as conformity assessment activities."

"specified requirement: need or expectation that is stated

- Note 1 to entry: Specified requirements can be stated in normative documents such as regulations, standards, and technical specifications.
- Note 2 to entry: Specified requirements can be detailed or general."

"object of conformity assessment: entity to which specified requirements apply Example: product, process, service, system, installation, project, data, design, material, claim, person, body or organization, or any combination thereof."

"conformity assessment body: body that performs conformity assessment activities, excluding accreditation"

"accreditation body: authoritative body that performs accreditation

- Note 1 to entry: The authority of an accreditation body can be derived from government, public authorities, contracts, market acceptance or scheme owners."

"first-party conformity assessment activity: conformity assessment activity that is performed by the person or organization that provides or that is the object of conformity assessment"

"second-party conformity assessment activity: conformity assessment activity that is performed by a person or organization that has a user interest in the object of conformity assessment"

"third-party conformity assessment activity: conformity assessment activity that is performed by a person or organization that is independent of the provider of the object of conformity assessment and has no user interest in the object"

"conformity assessment scheme/conformity assessment program: set of rules and procedures that describe the objects of conformity assessment, identify the specified requirements, and provide the methodology for performing conformity assessment"

"testing: determination of one or more characteristics of an object of conformity assessment, according to a procedure

- Note 1 to entry: The procedure can be intended to control variables within testing as a contribution to the accuracy or reliability of the results.
- Note 2 to entry: The results of testing can be expressed in terms of specified units or objective comparison with agreed references.
- Note 3 to entry: The output of testing can include comments (e.g., opinions and interpretations) about the test results and fulfillment of specified requirements."

"inspection: examination of an object of conformity assessment and determination of its conformity with detailed requirements or, on the basis of professional judgment, with general requirements"

"attestation: issue of a statement, based on a decision, that fulfillment of specified requirements has been demonstrated

- Note 1 to entry: The resulting (…) "statement of conformity", is intended to convey the assurance that the specified requirements have been fulfilled. Such an assurance does not, of itself, provide contractual or other legal guarantees.
- Note 2 to entry: First-party attestation and third-party attestation are distinguished by the terms declaration, certification, and accreditation, but there is no corresponding term applicable to second-party attestation."

"declaration: first-party attestation"

"certification: third-party attestation related to an object of conformity assessment, with the exception of accreditation"

"accreditation: third-party attestation related to a conformity assessment body, conveying formal demonstration of its competence, impartiality and consistent operation in performing specific conformity assessment activities"[19]

3.2 Towards a certification of trustworthy AI

Certification is a proven method to demonstrate the quality or desired properties of products, systems, or services for example. Based on an independent and expert attestation that specific requirements are being met, certificates provide an objective benchmark for third parties who themselves have no insight into the certified object. Thus, they are an important means to establish trust and, at the same time, increase competitiveness. Various prerequisites and steps prior to AI certification have been discussed in the previous chapters. **Section 1.1** examines the question of which requirements can be used to characterize the quality and trustworthiness of AI systems. **Section 1.2** discusses challenges regarding testing and verification of system requirements. Further, **Chapter 2** introduces a draft of a standard for AI management systems that describes those elements of an organization that should be implemented for the effective achievement of trustworthy AI. The questions that arise on the basis of these considerations are how the implementation of such a management system in accordance with requirements can be objectively confirmed and how requirements of trustworthiness in an AI system are certified. Certification schemes are needed in both cases. In particular, abstract requirements need to be translated to a list of measurable and observable criteria that can be checked against the respective object of certification and that ensures that all desired properties are present in it.

19 All quotations of terms and definitions are from [ISO/IEC, 2020b].

As discussed in **Chapter 2**, management systems, which are an established tool for organizations to effectively achieve their objectives, have been standardized and certified for decades. In addition, **Section 2.3** shows that the AIMS Draft describes a suitable framework for organizations to address AI risks as well as system requirements for trustworthy AI. As already mentioned, certification of AIMS would add value to organizations that want to generate evidence that they are well positioned to solve challenges in the face of AI and to manage specific risks. **Section 3.2.1** elaborates on the view that, given the long-standing approach to certifying management systems, it can be assumed that certification of AIMS can be implemented in a similar way.

However, the question of the certifiability of AI systems is more complex. IT systems that can modify their decision rules during operation, or whose behavior is to a large extent not predictable and hence requires systematic human oversight and risk assessment, have not been the object of certification to date. AI specifics, such as the potentially high degree of autonomy, must be taken into account accordingly when defining the scope of certification and respective certification schemes. Moreover, requirements for conformity assessment bodies and accreditation need to be defined. An interesting aspect here is the qualification of auditors who might need to be competent not only in the area of data science but also in engineering, management, and, potentially, to a certain extent, in a particular application domain like medicine. **Section 3.2.2** discusses steps that need to be taken towards AI certification and presents an approach to address AI-specific challenges when it comes to developing certification schemes.

3.2.1 Towards certification of AI management systems

As illustrated in the introduction and **Chapter 2**, certification of a management system (MS) generates evidence of the responsibility and accountability of an organization and can simplify compliance dependencies or regulatory obligations. Especially in view of the upcoming EU regulation, but also in light of AI incidents[20] that may weaken societal or customer's trust and thus competitiveness of the AI provider, a need has emerged for a certificate that attests to the trustworthy use and responsible management of AI-related matters.

Section 2.3 shows that the AIMS Draft outlines a reliable basic framework for governance, risk, and compliance (GRC) in the context of AI, with room for maneuver for organizations in the concrete design of their management processes as well as the technical-organizational measures for AI risk control. This result gives a sense that AIMS can help organizations to effectively achieve trustworthy use of AI. Furthermore, certification of AIMS would add value to organizations that want to demonstrate trustworthiness to third parties, such as (potential) customers or business partners. Consequently, once the standard is published, the necessary arrangements should be made to put certification with AIMS into practice.

Certification in general means that conformity with given requirements is attested by an independent party that has no particular user interest (see **Section 3.1**). The question of what requirements characterize such an independent party (conformity assessment body or certification body) is addressed by the ISO committee on conformity assessment (CASCO) which develops policy and publishes standards related to conformity assessment. Especially, CASCO has published the international standard "Conformity assessment - Requirements for bodies providing audit and certification of management systems" (ISO/IEC 17021-1:2015 [ISO/IEC, 2015a]) that formulates requirements concerning, amongst others, the competence, independence, and neutrality of certification bodies for management systems. Confirmation that a body in a defined scope meets (standardized) requirements and can competently perform certain conformity assessment tasks is called accreditation and is usually a sovereign task. Competent and neutral certification bodies will also be needed to issue certificates for AI management systems. In this regard, requirements for bodies that are supposed to certify AI management systems may be specified in addition to ISO/IEC 17021-1:2015. However, it can be assumed that accreditation requirements, as well as the scope and qualification of auditors for AIMS, can be set up and implemented in a similar way as is the case regarding certification for existing management system standards.

Another essential step to put certification based on AIMS into practice is to draw up a corresponding auditing and certification scheme. Such a scheme is intended to describe, among other things, by whom, under which conditions and according to which procedure or process AIMS is audited and, if successful, certified. Moreover, abstract requirements in AIMS need to be translated to a list of measurable and observable criteria that can be checked against and that ensures that all structures and processes that must be present in a particular AI management system are in fact implemented. Looking at the multitude of existing standards for management systems, auditing and certification schemes have already been established for many of them. In general, such schemes can be created, for instance, by governments, regulatory bodies, certification bodies, industry, or trade organizations. They are often derived not only from requirements formulated within the respective MS standard but

20 For an overview of AI incidents, see [McGregor, n.d.].

also from controls and implementation guidance given as recommendations in the standard.

An example of a management system standard (MSS) that is, similar to AIMS, aligned with the ISO high-level structure (HLS, see **Section 2.1**) and for which there are proven auditing and certification schemes, is ISO/IEC 27001:2013 [ISO/IEC, 2013], the international standard that specifies requirements for information security management system (ISMS). It belongs to a family of standards that, amongst others, comprises

- a standard that defines common terms and vocabulary related to information security management systems (ISO/IEC 27000:2018 [ISO/IEC, 2018a]),
- standards that provide implementation guidance regarding the requirements in ISO/IEC 27001:2013 (for instance ISO/IEC 27003:2017 [ISO/IEC, 2017a] and ISO/IEC 27005:2018 [ISO/IEC, 2018b]),
- a standard that provides guidelines for monitoring, measurement, analysis, and evaluation of an ISMS (ISO/IEC 27004:2016 [ISO/IEC, 2016]),
- a standard that, in addition to the requirements contained within ISO/IEC 17021-1 and ISO/IEC 27001, specifies requirements and provides guidance for bodies providing audit and certification of an ISMS (ISO/IEC 27006:2015 see [ISO/IEC, 2015c]),
- and further standards that contain requirements and guidance for sector-specific application of ISMSs (ISO/IEC 27009:2020 see [ISO/IEC, 2020b] and ISO/IEC 27010:2015 see [ISO/IEC, 2015d] to ISO/IEC 27019:2017 see [ISO/IEC, 2017b]).

Also, the German Federal Office for Information Security (BSI) has developed a holistic approach for implementing appropriate information security for organizations of all types and sizes. The methodology of this approach, known as "IT-Grundschutz"[21], is described in BSI standard 200-2 [BSI, 2017] and, together with BSI standard 200-3 on risk analysis and the "IT-Grundschutz Kompendium", which contains security requirements for different operational environments, forms a tool for selecting and adapting measures for the secure handling of information for an organization. Moreover, the BSI's "IT-Grundschutz" methodology describes a way to build an information security management system that is fundamentally compatible with the requirements in ISO/IEC 27001:2013. Especially, the BSI has derived a certification scheme from "IT-Grundschutz" [BSI, 2019] and acts as a certification body for ISMS.

The BSI certification scheme for ISMS describes the interplay between applicant, auditor and certification body and specifies the course and time frame of the certification procedure. The procedure starts when a certification request is submitted to the BSI. In this regard, the scheme stipulates which documents the applicant must submit with the request. Moreover, some requirements are intended to protect the confidentiality of reference documents. If the certification request is accepted, an audit of the applicant's ISMS is conducted, at the end of which an audit report is submitted to the certification body. Eventually, based on this audit report, the certification body decides whether a certificate will be issued or not. If successful, the certificate is valid for three years, provided that the applicant's ISMS passes annual control audits within that term.

From the description of the BSI's certification scheme for ISO/IEC 27001:2013, it becomes clear that audits, especially the one in the procedure for initial certification, have an essential influence on the awarding or retention of the certificate. Thus, the BSI, which is supposed to be independent and neutral in its function as a certification body, intends to ensure the comparability of audits and audit reports. Therefore, it has also developed an auditing scheme based on "IT-Grundschutz" [BSI, 2020b]. This stipulates, amongst others, that an audit consists of two phases. First, reference documents created by the applicant are reviewed. Second, an on-site audit verifies the practical implementation of the security measures documented in the reference documents for their adequacy, correctness, and effectiveness. To ensure the expertise, experience, and independence of auditors, it further requests that auditors are BSI-certified for auditing ISO/IEC 27001:2013 based on "IT-Grundschutz". Also, the format and content of the audit report, which the auditor must submit to the BSI, are defined in detail in the auditing scheme.[22]

The overview of the certification scheme for ISMS given in the previous paragraphs provides an impression of what procedures and higher-level requirements (e.g., for auditors) need to be defined and what else has to be set up (e.g., schemes and certification bodies) in order to put certification of AIMS into practice. It also becomes clear that for the development of certification schemes it is important that controls and recommendations for the implementation of an AI management system are of such detail and scope that criteria can be derived which permit a practical evaluation of any such system. Looking at the controls provided in AIMS, they seem to have an appropriate degree of detail so that certification schemes can be derived from it. Moreover, proven auditing and certification practices, undertaken in accordance with existing management

21 For more information, see [BSI, n. d. a].

22 The information on certification and auditing according to ISO 27001 based on "IT-Grundschutz", that is given in this and the previous paragraph, is provided in [BSI, 2019]. For more information, see also [BSI, n. d. b].

system standards like ISMS, yield a strong foundation for designing a certification for AIMS.

3.2.2 Towards certification of AI systems

The previous section elaborated on how organizations can make an essential contribution to safeguarding against AI risks and ensuring the quality of their AI products and services through organizational structures and processes. Especially, they can demonstrate trustworthiness through appropriate AI management. However, a central building block for quality and trust in AI systems still consists of implementing, proving, and eventually certifying technical system properties. Unlike AIMS, where it is perhaps possible to implement audits and certification in a similar manner to that for established MSSs like ISMS, demonstrating the desired properties of AI systems (see **Section 1.2.1**) presents a greater challenge. This section discusses what needs to be taken into account when developing certification schemes for AI systems. In particular, based on the considerations in the previous chapters, it becomes clear that system properties and their maintenance should not be tested and/or confirmed in isolation from the existence of corresponding organizational structures and processes. Moreover, the approach of the Fraunhofer IAIS audit catalog [Poretschkin, 2021] is presented, which follows the structuring of AI dimensions as elaborated in [Cremers, 2019]. It offers a possible procedure from which a certification or auditing scheme for AI systems can be derived. Finally, aspects of a certification infrastructure, which needs to be created to put certification of AI systems into practice, are considered.

As [Shneiderman, 2020] illustrates, contributions to the trustworthiness of AI can be made from different points of view. He distinguishes between "team, organization, and industry". While the team is responsible for verifying the development processes of the AI system and providing documentation trails, the organization should implement corresponding processes and manage resources and risks (which corresponds to the idea of an AI management system). At the industry level, a contribution to trustworthiness can be made by defining and implementing certification procedures for AI systems. In this regard, in recent years many companies but also states and non-governmental organizations have published AI ethics guidelines (see [Jobin, 2019]). Moreover, as broadly illustrated in previous chapters, regulators and standardization organizations (specifically the European Commission, but also national institutions such as the BSI in Germany) have shown intensified efforts to impose certification requirements on broad classes of AI systems. As can be seen from **Section 1.1** and **2.3**, some of these requirements need to be further operationalized to make them suitable for viable auditing and certification schemes. If possible, requirements should range all the way down to the application- or even algorithm level and capture the entire range of interconnectedness from standalone software products to cloud services. However, in many cases, the operationalization of requirements strongly depends on the specific application context. In other domains like IT security and functional safety, where a wide range of demands regarding the resistance to manipulation, resilience, or avoidance of unintentional misbehavior can result in quite diverse technical requirements for different systems, this challenge is addressed by risk-based testing and auditing approaches. These enable comparability between test results of diverse kinds of AI systems despite their highly different individual requirements. One concrete example of a risk-based approach is the methodology of "IT-Grundschutz" [BSI, 2017] on which the BSI's certification scheme for ISO/IEC 27001:2013 is based (see **Section 3.2.1**).

Following the Fraunhofer IAIS audit catalog, a risk-based audit and certification approach can be adapted to AI systems. The procedure presented in the Fraunhofer IAIS audit catalog is based on a detailed and systematic analysis of risks that is carried out separately for each technical dimension (see **Section 2.3.2**). The risk analysis is complemented by a documentation of measures that have been taken to mitigate the risks that are specific to the respective AI system to an acceptable level.

The procedure of the Fraunhofer IAIS audit catalog stipulates that first, a profile of the AI system is created as a starting- and reference point for the actual risk analysis in each technical dimension. The profile specifies the intended scope and functionality of the AI system as well as the limits of its application. Subsequently, a risk analysis is conducted which consists of two steps. In the first step, similar to "IT-Grundschutz" ([BSI, 2017]), the protection requirement is analyzed and categorized as low, intermediate, or high for each technical dimension. In the second step, those dimensions for which an intermediate or high protection requirement has been identified are examined in more detail. Each technical dimension consists of several so-called risk areas that correspond to the basic risk classes within that dimension. For each risk area within a relevant dimension, a detailed risk analysis is conducted that especially takes into account the specific application environment and context. Based on this risk analysis, requirements are formulated Agents involved in the development, deployment, or maintenance of the AI system choose and document effective measures for risk mitigation in accordance with these requirements. This two-step procedure bears similarities to the standardized risk management as detailed in ISO/IEC 31000:2018 (see also **Section 2.3.1.1**), where risk assessment and risk treatment are separate steps within an iterative process.

The broad applicability of the risk-based procedure of the Fraunhofer IAIS audit catalog prohibits quantitative specifications of minimum requirements and thresholds that might fit in one application but would be too weak or restrictive for

others. In addition, the wide spectrum of AI systems poses the challenge that both the possible hazards that are the cause of a risk and the measures for its mitigation cannot be presented in a complete and static list that a developer or auditor can refer to. The approach taken in the Fraunhofer IAIS audit catalog for addressing this inherent complexity is to let the AI system developer explicitly define and argue which goals should be achieved in the specific technical dimensions under consideration and how, for example, by which metrics the achievement of the goals should be measured. Application, documentation, and testing of risk mitigation measures then provide developers, auditors, and users with a practicable, usable scheme to substantially test and assess the risks and qualities of an AI system. The catalog provides a good orientation to most of the relevant state-of-the-art risk metrics and mitigation measures, which can be used as a reliable frame of reference.

Another important issue that should be taken into account when developing certification schemes for AI systems is that a lot of features that technically constitute a trustworthy AI system intrinsically require process regulation. This is already echoed in many of the trustworthiness requirements considered in **Section 2.3**, since AI-specific risks, for example those arising from continuous learning of the system or unpredictability of results, can often only be effectively controlled in conjunction with human oversight. In particular, the assessment of technical requirements and implementation details is always valid only as a snapshot. Hence, the actual trustworthiness of the AI system, especially one that continues to learn as it operates, remains questionable if its provider does not regulate and organize processes that guarantee both the adherence to specified requirements and a continuous and thorough monitoring of its operation after it has gone into production. Thus, auditing and certification schemes for AI systems cannot exclusively refer to the examination of the technical system itself but should also include the examination of relevant organizational structures and processes.

A recently proposed concept by [Zicari, 2021] for the inspection and certification of AI systems directly links the assessment of trustworthiness to the processes that are implemented in an organization. A more separating view in this respect is taken by the European Commission which requires that trustworthy AI shall be ensured through technical requirements for the AI system on the one hand and management-related obligations (in particular, risk management, post-market monitoring, and quality management) on the other. Regarding the former, [Mock et. al., 2021], for example, discuss the (technical) product-related view on trustworthy AI using the use case of autonomous driving. Also, the Fraunhofer IAIS audit catalog primarily takes a product view and provides a highly suitable procedure to address the technical requirements in the proposed AI regulation. In part, the audit catalog addresses relevant organizational processes within its so-called risk areas "control of dynamics", which focus on the monitoring and maintenance of desired system properties and risk mitigation. However, the audit catalog does not look at organizations as comprehensively as AIMS does, in the sense that it would, for example, consider the planning of resources. Regarding the process-related requirements in the proposed regulation on AI, it is precisely via properly designed management systems that they can be addressed. Providing international standards for the certification of these processes adds an important part on the way to AI certification.

When developing a certification scheme, the validity period of certificates must also be regulated. The fact that product tests or audits always represent only a snapshot of the system, cannot be easily solved. Approaches to continuous certification, which are being researched in the cloud area, for example, have not yet been established. Thus, given the dynamic nature of AI systems, regular control audits after initial certification, such as those required by the BSI's certification scheme based on "IT-Grundschutz", appear to be appropriate. Ensuring that processes and responsibilities for monitoring and maintaining system properties exist within an organization should also have an impact on the validity period of product certificates, or even be a prerequisite for product certification in the case of particularly critical applications. One possible approach is to link the validity of product certificates to the certification of AIMS. If not being a set prerequisite, this could take the form, for example, of increasing the intervals of control audits of the AI system through AIMS certification.

In addition to describing the specific testing, auditing, and certification procedure, a workable certification scheme should also define requirements for auditors or assessors, for example in terms of their qualifications. Here it should be taken into account that risks, especially for data-driven technologies such as ML, strongly depend on the application context and the specific domain. Consequently, it can be assumed that in many cases domain-specific expertise is needed to derive requirements for an AI system and evaluate it. Given the consideration that auditing and certification of AI require examination of the system itself as well as, in part, processes within the organization, and, in addition, domain expertise may be required, the qualifications of auditors should be appropriately broad. Another approach is to have interdisciplinary teams conduct the audit of an AI system. For example, testing of an application for image recognition in the medical field could be performed by an auditor with technical expertise in data science and engineering, together with an auditor with expertise in management and a radiologist.

Moreover, accreditation requirements for certification bodies need to be defined that address the specific challenges of AI. The international standard ISO/IEC 17065:2012 "Conformity assessment – Requirements for bodies certifying products,

processes, and services" [ISO/IEC, 2012] may serve as a guideline and be developed with regard to AI specifics. In particular, an infrastructure for testing AI systems needs to be further developed. As described in **Section 1.2.1**, testing and verifying system-related requirements is an active area of research that has not yet provided for comprehensive sets of AI testing tools. Thus, another step that needs to be taken to put certification of AI systems into practice is to establish test laboratories with (potentially standardized) testing tools that may become part of auditing schemes. Another challenge that needs to be addressed is the degree to which auditors are allowed to view technical details of the system and (confidential) training data and operational data in particular. In addition, operators often rely on external components, especially cloud services, due to the high computing power and data volumes required in the face of ML technologies. Thus, it is also necessary to regulate how insight into third-party components that are integrated into an AI system or traceability of their development process may not be possible.

In summary, the procedure suggested by the Fraunhofer IAIS audit catalog gives an idea that risk-based testing and auditing can constitute the central component of upcoming certification procedures for AI systems. Compared to the field of management systems (see **Section 3.2.1**), what is still missing in broad areas for putting audits and certification of AI systems into practice, are globally accepted standards that could provide a reliable and binding framework. Catalogs such as the Fraunhofer IAIS audit catalog or the BSI AIC4 catalog have the potential to fill this gap, as they allow a substantial, practical, and comparable risk assessment and can be employed by agents of all levels of the certification hierarchy, from developers and users of AI systems to regulatory- and certification bodies. In addition to defining an audit and certification scheme for AI systems, a corresponding infrastructure needs to be created that is suitable to put these schemes into practice. Here, defining the qualification of auditors and establishing test laboratories and AI testing tools are key steps towards implementing audits and certification of AI systems.

Conclusions and Recommendations for Action

It is clear from the detailed analyses performed in this study that the AIMS draft is an important and adequate step towards achieving and supporting the trustworthy development and use of AI technologies in companies and organizations. The comparison with the HLEG requirements, the proposed EU regulation on AI, and the BSI AIC4 standard (denoted as "other documents" for short) in the following leads in particular to the following conclusions:

1. Clear distinction between organizational and product perspective

The interplay between these two perspectives has been elaborated in **Section 1.2**. Both perspectives need to be addressed for achieving trustworthy AI and are often implicitly mixed. The analysis in **Section 2.3.1** reveals that the AIMS draft is much more concise and advanced in formulating organizational requirements than the "other documents", in particular regarding accountability and resources. We would recommend emphasizing this advantage of the AIMS Draft while making clear the more product-oriented issues are set out in detail in other upcoming standards of the ISO/IEC working group SC 42.

2. Coverage of technical dimensions

The detailed analysis carried out in **Section 2.3.2** showed that almost all the technical requirements set out in detail in the "other documents" are also addressed in the AIMS draft. Although all relevant dimensions are clearly covered, we would recommend checking again whether individual aspects such as "uncertainty", "detection of malicious inputs", and "record keeping" should achieve more explicit attention in the AIMS draft

3. Definition of terminology; risk

One major approach followed and agreed by the "other documents", as well as the state-of-the-art in trustworthy, AI is commonly denoted as a "risk-based" approach. For example, the proposed EU regulation on AI explicitly requires the establishment of a risk management system for high-risk AI applications. Here, potential risks emerging from the AI system are meant. The AIMS draft also requires and emphasizes risk management. It differentiates between risks for the organization and the potential impact of the AI system. While the "other documents" subsume these two aspects under the general notion of risks, the AIMS Draft requires the undertaking of an AI-specific impact assessment that addresses the same AI issues denoted as "risks" in the "other documents." This leaves the decision open to the company or organization, whether the required impact assessment is integrated into existing risk management processes or whether it is performed explicitly. We recommend mentioning explicitly the relationship between "impact assessment" as required by the AIMS draft and the "risk management" as required by the "other documents".

4. Towards certification

The analysis in **Chapter 3** has shown that the AIMS standard provides a sufficient level of coverage and detail such that, for example, regulatory- and certification bodies can derive practical certification schemes for the certification of AI management systems. However, even if all relevant technical dimensions a covered, the level of detail that would be needed for a product level certification can and should not be achieved in a management system standard (see discussion about the Fraunhofer IAIS audit catalog for trustworthy AI in **Section 3.2**). Regarding the certification of AI management systems, we recommend actively seeking interaction with regulatory bodies or certification bodies to develop a certification scheme.

5. EU requirement to establish a "Quality Management System"

The proposed EU regulation on AI explicitly requires that a "Quality Management System" should be established for high-risk AI applications. Risk management and data governance over the whole life cycle of an AI application should be integral parts of the "Quality Management System". Our analysis in **Chapter 2** shows that EU requirements on such a "Quality Management System" are also addressed in the AIMS draft, although the AIMS draft does not refer to "Quality Management System". We recommend seeking the discussion and clarification with the regulatory authorities regarding this point.

6. Broaden the scope of overall goals

The comparison of the AIMS draft with the "other documents" has revealed a great harmony in the general understanding of what the goals and rationales of "Trustworthy AI" should be. Nevertheless, ISO/IEC is acting in a worldwide context, not just limited to Europe or Germany. Hence, it may be worthwhile to also consider overall goals that can be accepted worldwide, such as the sustainability goals of the United Nations, when defining motivations and goals of the AIMS standard. Similarly, it is to be expected that the European regulation on data protection (GDPR) will have a large impact in the context of the upcoming "enforcement" of the proposed regulation on AI. A positioning of or reference in the AIMS might also have a clarifying effect.

Overall, it is a most reasonable approach to anchor the task of ensuring a trustworthy development and use of AI technologies in the GRC strategy of companies, as in the fields of privacy and security that have already been addressed. The implementation of management systems that support and incorporate AI-specific requirements will significantly contribute to customers' trust and make it easier to achieve compliance over the whole life-cycle of AI systems, even in the presence of multiple stakeholders and complex supply chains. The AIMS draft is a solid and valid foundation of such management systems.

References

[Antunes, 2018] Antunes, N., et al., Fairness and transparency of machine learning for trustworthy cloud services. 2018 48th Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W). IEEE, 2018.

[Arnold, 2019] Arnold et al., FactSheets: Increasing Trust in AI Services through Supplier's Declarations of Conformity, in: IBM Journal of Research and Development, 2019.

[Arya, 2020] Arya et al., AI Explainability 360: An Extensible Toolkit for Understanding Data and Machine Learning Models, in: Journal of Machine Learning Research, 2020.

[Askell, 2019] Askell et.al., The Role of Cooperation in Responsible AI Development. arXiv: 1907.04534, 2019. URL: http://arxiv.org/abs/1907.04534 (Accessed: 11.08.21)

[Bellamy, 2019] Bellamy et al., AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Un-wanted Algorithmic Bias, in: IBM Journal of Research and Development, 2019.

[Brundage, 2020] Brundage, M. et al., Toward trustworthy AI development: mechanisms for supporting verifiable claims. arXiv preprint arXiv:2004.07213, 2020.

[BSI, 2017] German Federal Office for Information Security (BSI), BSI-Standard 200-2. IT-Grundschutz-Methodik, 2017. URL: https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/Grundschutz/BSI_Standards/standard_200_2.html?nn=128640 (Accessed: 11.08.2021)

[BSI, 2019] German Federal Office for Information Security (BSI), Zertifizierung nach ISO 27001 auf der Basis von IT-Grundschutz - Zertifizierungsschema; Version 2.1, 2019. URL: https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/Grundschutz/Zertifikat/ISO27001/Zertifizierungsschema_Kompendium.html (Accessed: 11.08.2021)

[BSI, 2020a] German Federal Office for Information Security (BSI), Cloud Computing Compliance Criteria Catalogue – C5:2020, 2020. URL: https://www.bsi.bund.de/EN/Topics/CloudComputing/Compliance_Criteria_Catalogue/C5_NewRelease/C5_NewRelease_node.html (Accessed: 15.09.2021)

[BSI, 2020b] German Federal Office for Information Security (BSI), Zertifizierung nach ISO 27001 auf der Basis von IT-Grundschutz - Auditierungsschema; Version 2.3, 2020. URL: https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/Grundschutz/Zertifikat/ISO27001/Auditierungsschema_Kompendium.html (Accessed: 11.08.2021)

[BSI, 2021] German Federal Office for Information Security (BSI), AI Cloud Service Compliance Criteria Catalogue (AIC4), 2021. URL:https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/CloudComputing/AIC4/AI-Cloud-Service-Compliance-Criteria-Catalogue_AIC4.html (Accessed: 07.07.2021)

[BSI, n. d.a] German Federal Office for Information Security (BSI). IT-Grundschutz: Informationssicherheit mit System, n. d. URL: https://www.bsi.bund.de/DE/Themen/Unternehmen-und-Organisationen/Standards-und-Zertifizierung/IT-Grundschutz/it-grundschutz_node.html (Accessed: 15.09.2021)

[BSI, n. d.b] German Federal Office for Information Security (BSI). Zertifizierungsschema, n. d. URL: https://www.bsi.bund.de/DE/Themen/Unternehmen-und-Organisationen/Standards-und-Zertifizierung/Zertifizierung-und-Anerkennung/Zertifizierung-von-Managementsystemen/ISO-27001-Basis-IT-Grundschutz/Zertifizierungsschema/schema_node.html (Accessed: 15.09.2021)

[Burton, 2017] Burton et al., Making the Case for Safety of Machine Learning in Highly Automated Driving, in: Tonetta et al. (Hrsg.), Computer Safety, Reliability, and Security. SAFECOMP, 2017.

[Chiregi, 2018] Chiregi, et al., Cloud computing and trust evaluation: A systematic literature review of the state-of-the-art mechanisms. Journal of Electrical Systems and Information Technology 5.3, 2018.

[Coeckelbergh, 2020] Coeckelbergh, M., Artificial intelligence, responsibility attribution, and a relational justification of explainability. Science and engineering ethics 26.4, 2020.

[Cremers et. al. 2019] Cremers, Englander et al., Trustworthy use of Artificial Intelligence: Priorities from a philosophical, ethical, legal, and technological viewpoint as a basis for certification of Artificial Intelligence. Fraunhofer IAIS, 2019 URL: https://www.iais.fraunhofer.de/content/dam/iais/KINRW/Whitepaper_Thrustworthy_AI.pdf, (Accessed: 11.08.2021)

[DIN e.V. & DKE, 2020] DIN e.V. & DKE, Deutsche Normungsroadmap Künstliche Intelligenz, 2020. URL: https://www.din.de/resource/blob/772438/6b5ac6680543eff9fe372603514be3e6/normungsroadmap-ki-data.pdf (Accessed: 15.09.2021)

[EC, 2021] European Commission, Proposal for a Regulation of the European Parliament and of the Council laying down Harmonized Rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts., 2021. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX-:52021PC0206 (Accessed: 15.08.2021)

[Floridi, 2016] Floridi et al., What is data ethics? in Philosophical Transactions of the Royal Society: Mathematical Physical and Engineering Sciences, 2016.

[Floridi, 2018] Floridi et al., AI4People - An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations in Minds & Machines, 2018.

[Gebru, 2018] Gebru et al., Data Sheets for Datasets, in: Proceedings of the 5th Workshop on Fairness, Accountability, and Transparency in Machine Learning, PLMR 80, 2018.

[Hallensleben, 2020] Hallensleben et al., From Principles to Practice – An interdisciplinary framework to operationalize AI ethics, Bertelsmann Stiftung, 2020.

[Hess, 2018] Hess et al., The Trustworthy Pal: Controlling the False Discovery Rate in Boolean Matrix Factorization, in: Proceedings of the 2018 SIAM International Conference on Data Mining, 2018.

[HLEG, 2019a] High-level Expert Group on AI , Policy and investment recommendations for trustworthy Artificial Intelligence, 2019. URL: https://digital-strategy.ec.europa.eu/en/library/policy-and-investment-recommendations-trustworthy-artificial-intelligence (Accessed. 15.09.2021)

[HLEG, 2019b] High-level Expert Group on AI, Ethics Guidelines on Trustworthy AI, 2019.

[HLEG, 2020] High-level Expert Group on AI, The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self assessment, 2020.

[Hodges, 2016] Hodges, C., Ethical business regulation: understanding the evidence. Birmingham: Department for Business Innovation & Skills, Better Regulation Delivery Office, 2016. URL: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/497539/16-113-ethical-business-regulation.pdf (Accessed: 15.08.2021)

[Huang, 2017] Huang et al., Safety Verification of Deep Neural Networks, in Kunčak und Majumdar (Hrsg.), Computer Aided Verification. CAV, 2017.

[ISO/IEC, 2012] International Organization for Standardization. Standard ISO/IEC 17065:2012. Conformity assessment – Requirements for bodies certifying products, processes and services, 2012.

[ISO/IEC, 2013] International Organization for Standardization. Standard ISO/IEC 27001:2013. Information technology – Security techniques – Information security management systems — Requirements, 2013.

[ISO/IEC, 2015a] International Organization for Standardization. Standard ISO/IEC 17021-1:2015. Conformity assessment – Requirements for bodies providing audit and certification of management systems – Part 1: Requirements, 2015.

[ISO/IEC, 2015b] International Organization for Standardization. Standard ISO/IEC 27001:2013/Cor 2:2015. Information technology – Security techniques – Information security management systems – Requirements – Technical Corrigendum 2, 2015.

[ISO/IEC, 2015c] International Organization for Standardization. Standard ISO/IEC 27006:2015. Information technology – Security techniques – Requirements for bodies providing audit and certification of information security management systems, 2015.

[ISO/IEC, 2015d] International Organization for Standardization. Standard ISO/IEC 27010:2015. Information technology – Security techniques – Information security management for inter-sector and inter-organizational communications, 2015.

[ISO/IEC, 2016] International Organization for Standardization. Standard ISO/IEC 27004:2016. Information technology – Security techniques – Information security management – Monitoring, measurement, analysis and evaluation, 2016.

[ISO/IEC, 2017a] International Organization for Standardization. Standard ISO/IEC 27003:2017. Information technology – Security techniques – Information security management systems – Guidance, 2017.

[ISO/IEC, 2017b] International Organization for Standardization. Standard ISO/IEC 27019:2017. Information technology – Security techniques – Information security controls for the energy utility industry, 2017.

[ISO, 2018] International Organization for Standardization. Standard ISO 31000:2018. Risk management – Guidelines, 2018.

[ISO/IEC, 2018a] International Organization for Standardization. Standard ISO/IEC 27000:2018. Information technology – Security techniques – Information security management systems – Overview and vocabulary, 2018.

[ISO/IEC, 2018b] International Organization for Standardization. Standard ISO/IEC 27005:2018. Information technology – Security techniques – Information security risk management, 2018.

[ISO/IEC, 2020a] International Organization for Standardization. Standard ISO/IEC 17000:2020. Conformity assessment – Vocabulary and general principles, 2020.

[ISO/IEC, 2020b] International Organization for Standardization. Standard ISO/IEC 27009:2020. Information security, cybersecurity and privacy protection – Sector-specific application of ISO/IEC 27001 – Requirements, 2020.

[ISO/IEC, 2021] ISO/IEC Directives, Part 1, Consolidated ISO Supplement, 2021. URL: https://isotc.iso.org/livelink/livelink/fetch/2000/2122/4230450/4230452/Consolidated_ISO-IEC_Part-1_%28E%29_2021.pdf?nodeid=21825221&vernum=-2 (Accessed: 15.09.2021); Annex SL Appendix 2: (normative) Harmonized structure for MSS with guidance for use. URL: https://isotc.iso.org/livelink/livelink/fetch/-8921878/8921901/16347356/16347818/2021-05_Annex_SL_Appendix_2.pdf?nodeid=21826538&vernum=-2 (Accessed: 15.09.2021)

[ISO/IEC, 2021b] International Organization for Standardization. Standard ISO/IEC CD 23894. Information Technology – Artificial intelligence – Risk Management, under development.

[ISO/IEC, 2021c] International Organization for Standardization. Standard ISO/IEC CD 42001. Information Technology – Artificial intelligence – Management system, under development.

[ISO/IEC, 2021d] International Organization for Standardization. Standard ISO/IEC DIS 38507. Information technology – Governance of IT – Governance implications of the use of artificial intelligence by organizations, under development.

[ISO, 2021] ISO. Management System Standards, n.d. URL: https://www.iso.org/management-system-standards.html

[Jobin, 2019] Jobin et al., The global landscape of AI ethics guidelines in Nature Machine Intelligence, 2019.

[Kersten, 2020] Kersten, H. et al., IT-Sicherheitsmanagement nach der neuen ISO 27001. Springer Vieweg, 2020. ISBN 978-3-658-27691-1

[Lambert, 2017] Lambert, G., A stroll down Quality Street in ISOfocus 123 July-August 2017. pp. 37-40, 2017. URL: https://www.iso.org/files/live/sites/isoorg/files/news/magazine/ISOfocus%20(2013-NOW)/en/2017/ISOfocus_123/ISOfocus_123_EN.pdf

[Madaio, 2020] Madaio, M.et al., Co-designing checklists to understand organizational challenges and opportunities around fairness in ai. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 2020.

[McGregor, n.d.] Mc Gregor S., Preventing Repeated Real World AI Failures by Cataloging Incidents: The AI Incident Database, n.d. URL: https://incidentdatabase.ai/discover (Accessed:15.09.2021)

[Mitchell, 2019] Mitchell et al., Model Cards for Model Reporting, in: FAT* '19: Proceedings of the Conference on Fairness, Accountability, and Transparency, 2019.

[Mittelstadt, 2019] Mittelstadt, Brent. Principles alone cannot guarantee ethical AI. Nature Machine Intelligence 1.11 pp. 501-507, 2019. URL: https://arxiv.org/ftp/arxiv/papers/1906/1906.06668.pdf (Accessed: 15.08.2021)

[Mock et. al., 2021] Mock, M., Scholz, S.et al. An Integrated Approach to a Safety Argumentation for AI-based Perception Functions in Automated Driving, Workshop on Artificial Intelligence Safety Engineering at Safecomp 2021, York, 2021 (2021)

[Nicolae, 2018] Nicolae et al., Adversarial Robustness Toolbox, 2018. URL: https://arxiv.org/abs/1807.01069 (Accessed: 15.08.2021)

[Nori, 2019] Nori, Harsha, et al. Interpretml: A unified framework for machine learning interpretability. arXiv preprint arXiv:1909.09223, 2019.

[Papineni, 2002] Papineni et al., Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th annual meeting on association for computational linguistics, 2002.

[Peters, 2020] Peters, Dorian, et al., Responsible AI—two frameworks for ethical design practice. IEEE Transactions on Technology and Society 1.1 pp. 34-47, 2020. URL: https://spiral.imperial.ac.uk/bitstream/10044/1/77602/5/09001063.pdf (Accessed: 15.08.2021)

[Poretschkin, 2021] Poretschkin, M., et al., KI-Prüfkatalog: Ein Leitfaden zur Gestaltung vertrauenswürdiger Künstlicher Intelligenz, Fraunhofer IAIS, 2021. URL: https://www.iais.fraunhofer.de/content/dam/iais/fb/Kuenstliche_intelligenz/ki-pruefkatalog/202107_KI-Pruefkatalog.pdf (Accessed: 11.08.2021)

[Saleiro, 2018] Saleiro et al., Aequitas: A Bias and Fairness Audit Toolkit, 2018. URL: https://arxiv.org/abs/1811.05577 (Accessed:11.08.2021)

[Salimans, 2016] Salimans et al., Improved techniques for training gans, in: Advances in Neural Information Processing Systems, 2016.

[Santos, 2017] Santos, W. d.; Carvalho, L. F. M. et al., Lemonade: A scalable and efficient spark-based platform for data analytics, in Proceedings of the 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, ser. CCGrid, 2017.

[Shneiderman, 2020] Shneiderman, B., Bridging the gap between ethics and practice: Guidelines for reliable, safe, and trustworthy Human-Centered AI systems. ACM Transactions on Interactive Intelligent Systems (TiiS) 10.4, 2020.

[Thiebes, 2020] Thiebes et al., Trustworthy artificial intelligence. in Electron Markets, 2020.

[Toreini, 2019] Toreini et al., The relationship between trust in AI and trustworthy machine learning technologies. In: arXiv arXiv: 1912.00782, 2019. URL: http://arxiv.org/abs/1912.00782 (Accessed: 15.08.2021)

[Verma, 2018] Verma und Rubin, Fairness definitions explained, in: 2018 IEEE/ACM International Workshop on Software Fairness (FairWare), 2018.

[Weng, 2018] Weng et al., Evaluating the Robustness of Neural Networks: An Extreme Value Theory Approach, in: Sixth International Conference on Learning Representations, 2018.

[Zhao, 2020] Xingyu Zhao et al., A Safety Framework for Critical Systems Utilising Deep Neural Networks.In: arXiv, 2020 URL: https://arxiv.org/abs/2003.05311 (Accessed: 15.08.2021)

[Zicari, 2021] Zicari, Roberto V., et al.,Z-Inspection®: A Process to Assess Trustworthy AI. IEEE Transactions on Technology and Society, 2021.

[Zweig, 2018] K. A. Zweig et al., Wo Maschinen irren können. Verantwortlichkeiten und Fehlerquellen in Prozessen algorithmischer Entscheidungsfindung. Gütersloh: Bertelsmann Stiftung, 2018. URL: https://www.bertelsmann-stiftung.de/fileadmin/files/BSt/Publikationen/GrauePublikationen/WoMaschinenIrren-Koennen.pdf (Accessed: 15.08.2021)

Imprint

Publisher
Fraunhofer Institute for Intelligent Analysis
and Information Systems IAIS
Schloss Birlinghoven 1
53757 Sankt Augustin

Editors
Daria Tomala
Silke Loh

Layout and design
Achim Kapusta
Angelina Lindenbeck

Image sources
Titelbild: © Alex - stock.adobe.com/Fraunhofer IAIS

Date
October 2021

Contact

Fraunhofer Institute for Intelligent Analysis
and Information Systems IAIS
Schloss Birlinghoven 1
53757 Sankt Augustin

www.iais.fraunhofer.de

PD. Dr. Michael Mock
michael.mock@iais.fraunhofer.de

Dr. Maximilian Poretschkin
maximilian.poretschkin@iais.fraunhofer.de