DIN DKE

GERMAN STANDARDIZATION ROADMAP
ON ARTIFICIAL INTELLIGENCE

2ND EDITION

Supported by:

Federal Ministry
for Economic Affairs
and Climate Action

on the basis of a decision
by the German Bundestag

PUBLISHED BY
Wolfgang Wahlster
Christoph Winterhalter

DIN e. V.

Am DIN-Platz
Burggrafenstr. 6
10787 Berlin
Tel.: +49 30 2601-0
Email: presse@din.de
Website: www.din.de

DKE German Commission for Electrical, Electronic & Information Technologies of DIN and VDE
Merianstraße 28
63069 Offenbach am Main
Tel.: +49 69 6308-0
Fax: +49 69 6308-9863
Email: dke@vde.com
Website: www.dke.de

Referencing of the German Standardization Roadmap on Artificial Intelligence:
DIN, DKE (2022): German Standardization Roadmap
on Artificial Intelligence (2nd edition);
www.din.de/go/roadmap-ai

Photo credit:
Cover: pinkeyes – stock.adobe.com
Chapter entry pages:
kras99 (S. 11, 51), assistant (S. 31), Thitichaya (S. 37),
Maxim (S. 53), peshkov (S. 99), gunayaliyeva (S. 121),
pickup (S. 145, 231, 255, 275, 285), LuckyStep (S. 167),
kaptn (S. 187), ryzhi (S. 211), Alex (S. 265) – stock.adobe.com

December 2022

FOREWORD

Prof. Dr. rer. nat. Dr. h.c. mult.
Wolfgang Wahlster
CEA of the German Research
Center for Artificial Intelligence
(DFKI)

Christoph Winterhalter
Chairman of the Executive
Board, DIN

Dear Reader,

With the second edition of the German Standardization Roadmap Artificial Intelligence, today we are able to present an expanded and updated analysis of the current status and need for international standards and specifications for this key technology. In doing so, we wish to build on the great success of the Roadmap published in November 2020, the results of which received a great amount of international attention not only in specialist circles but also among political bodies and by the press following the presentation of the Roadmap at the German government's Digital Summit.

Standardization is also seen by the new German federal government as part of its AI Strategy and as a topic of very great importance. Following a kick-off event organized by DIN and DKE on 20.01.2022, this new edition of the Standardization Roadmap AI was developed with the active participation of more than 570 experts from industry, science, civil society and politics in nine working groups. These were accompanied by a high-level coordination group on AI standardization and conformity with a mandate from the federal government. The work focused on nine core topics (basic topics, security/safety, testing/certification, sociotechnical systems, medicine, industrial automation, mobility, energy/environment, financial services), with AI topics in sociotechnical systems, energy and environment, and financial services being new topics. The comprehensive Final Report of the Enquete Commission AI of the Bundestag of October 2020 was also taken into account.

As part of the German government's AI Strategy, six AI competence centres have now been established as the core of the German AI research landscape. In addition to the further development of the German Research Centre for Artificial Intelligence (DFKI), five other AI competence centres located at universities were established and expanded, and from the middle of this year, they were given permanent institutional funding. Together with 100 newly filled AI professorships, this results in a major boost for AI research and transfer to industry, to which the implementation of the present Standardization Roadmap AI can make a significant contribution. For the much-needed skilled workforce, initiatives such as the AI Campus for online training and the AI Grid for the networking and mentoring of researching AI young talent in international micro-subject communities have been launched, among others. The Platform Learning Systems has been further developed into a central AI platform for dialogues among science, industry, society and politics, and is now also responsible for the regular monitoring of the implementation of the German AI Strategy using current key figures.

The German government's AI Strategy has ushered in a “golden age” for German researchers that is the envy of many colleagues in other countries. Even the U.S. and China governments hardly provide as many long-term funding opportunities for AI relative to population size as in Germany.

However, the task now is to reap the rewards of these government investments through increased transfer to economic value creation, including in small and medium-sized enterprises (SMEs) and through start-ups. In this context, the implementation phase of the Standardization Roadmap AI now starting in 2023/2024 is of utmost importance, because standards and the certification of AI solutions based on standards enable companies to achieve investment security, legal certainty, and interoperability between platforms and value networks, and secure market shares.

Rising energy costs and the threat of supply shortages, high inflation rates, and disrupted supply chains as a result of the Corona pandemic and the Russian war against Ukraine are putting pressure on the entire economy. Artificial intelligence, as a technology of the future in this difficult geopolitical situation, can quickly contribute partial solutions to overcome these severe economic and societal pressures. For this to happen, however, access to the data necessary for the use of AI technologies must be facilitated and not blocked in a way that hinders innovation. Standardization can also objectify the often critical public debate on AI risks, so that the necessary investments in this future technology are increasingly made by decision-makers in companies.

For Germany, the world's most innovative factory supplier, industrial AI plays a special role. Industrie 4.0 is an export hit, as the Hanover Fair has shown again this year. Industrial AI plays a significant role in this area. It forms the basis for implementing the fourth industrial revolution in adaptable, cyber-physical factories for small batch sizes, in which collaborative and cognitive robots work hand in hand with specialists in Al-based zero-defect production to produce high-quality, high-tech products in a climate-friendly manner. Digital twins measure the CO_2 footprint during production and use AI algorithms to help reduce energy consumption.

Today, Germany has the highest robot density in Europe, and the first collaborative robots were developed to commercial maturity in Germany. Currently, there are more manufacturers or their research labs for collaborative cognitive robots here than in other parts of the world. Germany is still the clear leader in the field of industrial AI, as our colleagues from the USA and China also confirm. However, the task now is to transfer this lead into industrial practice and to secure it in the long term through standards and specifications.

Of course, there are quite a few AI application areas where other nations are currently leading – but mostly for a very good reason: For example, Germany is not working on AI surveillance of civilians. Research on AI for personalized and ubiquitous advertising on the Internet or for the production of fully autonomous weapons systems is also not desired in Germany and is not funded by the state.

The timely preparation of this second edition of the Standardization Roadmap AI would not have been possible without the tireless efforts of our volunteer experts. The coordination group on AI standardization and conformity, consisting of high-ranking personalities, has held regular meetings and additionally adopted the final recommendations for action by consensus in a closed meeting together with the leaders of the nine working groups.

On behalf of the steering group, we would also like to take this opportunity to thank all active members of the working groups for their great commitment, and we would like to express our special praise to Ms. Filiz Elmas as the excellent coordinator of the overall project and head of her team.

Standardization in Germany is a joint task based on the broad participation and collaboration of experts from industry, science, government and society. Only the early engagement of experts with broad experience will make it possible to develop standards and specifications for AI that are in line with the market and with needs, and to ensure the acceptance of these standards and specifications. If Germany wants to ensure that its interests are adequately taken into account in international AI standards, it is necessary to integrate AI experts into our standardization bodies and to strengthen the participation of German experts in international AI standardization bodies.

We wish all readers an exciting read and ask for your active support in the implementation of this standardization roadmap over the coming years.

Together, we can then master quite a few challenges of these changing times through a targeted use of certified AI technologies in accordance with our European value system.

W. Wahlster

Prof. Dr. rer. nat. Dr. h.c. mult. Wolfgang Wahlster,
CEA of the German Research Center for Artificial Intelligence (DFKI)

Christoph Winterhalter
Chairman of the Executive Board, DIN

GREETING

Dr. Robert Habeck
German Federal Minister for Economic Affairs
and Climate Action

Dear Reader,

We associate the term “artificial intelligence” (AI) with high hopes for a future technology that can facilitate production, work and administrative processes. At the same time, there are also serious reservations, for example regarding the control and protection of civil rights. The task now is to minimize the risks in order to make the best possible use of the many opportunities that AI opens up for us.

For this, we need a legal framework that promotes the responsible development of AI technologies and their use, oriented toward humans and the common good, and that also creates trust among users.

We want to create such a legal framework with the Artificial Intelligence Act, which is currently being negotiated at EU level. The aim is to ensure that AI systems placed on the European single market – depending on their risk level – meet certain requirements for transparency, accuracy, IT security and human oversight, for example. The concrete technical implementation of this Act will then be specified in European Standards.

In parallel with the European processes, AI standards are also being developed at international level to define the future global market environment for this technology. Those who want to be at the forefront of the global competition for the best AI solutions need to help shape this market environment. We therefore stipulated in Germany’s Coalition Agreement that we want to strengthen our involvement in international standardization processes. By actively contributing to the development of international AI standards and specifications, we can pave the way for our innovations to enter global markets and ensure that European values are taken into account when developing and marketing new technologies.

In the second edition of the German Standardization Roadmap Artificial Intelligence, you will find an overview of the existing standardization landscape, concrete standardization needs that we from Germany and Europe can contribute to international standardization, and recommendations for action for the topics that need to be prioritized. The German Standardization Roadmap AI is an important basis for promoting “Artificial Intelligence (AI) Made in Germany” as a globally recognized seal of approval for a trustworthy technology. I would like to thank everyone involved at DIN and DKE for their excellent work.

The next step is to implement the identified recommendations for action. All stakeholders involved in standardization are called upon here. Get involved and seize the opportunity to help shape the rules of the game for artificial intelligence.

Robert Habeck

Your
Dr. Robert Habeck
German Federal Minister for Economic Affairs
and Climate Action

Summary

On behalf of the German Federal Ministry of Economic Affairs and Climate Action, DIN and DKE started work on the second edition of the German Standardization Roadmap Artificial Intelligence in January 2022. With the broad participation and involvement of more than 570 experts from industry, science, the public sector and civil society, the strategic Roadmap for AI standardization was thus further developed. This work was coordinated and accompanied by a high-level coordination group for AI standardization and conformity.

The standardization roadmap implements a measure of the German government's AI Strategy and thus makes a significant contribution to "AI – Made in Germany".

Standardization is part of the AI Strategy and is a strategic instrument for strengthening the innovation and competitiveness of the German and European economies. Not least for this reason, standardization plays a special role in the planned European legal framework for AI, the Artificial Intelligence Act.

This Standardization Roadmap AI identifies the requirements in standardization, formulates concrete recommendations and thus creates the basis for initiating standardization work at national level, and especially at European and international level, at an early stage. In doing so, the Roadmap makes a significant contribution to the European Commission's Artificial Intelligence Act, supporting its implementation

Chapter 1 of the Roadmap introduces the topic and presents the significance of standardization for economic policy, as well as the Roadmap's objectives and approach.

The current actors and standardization environment for AI are described in Chapter 3, which gives an overview of relevant innovation policy initiatives, research projects, and standardization activities.

The Standardization Roadmap AI focuses on nine key topics, which are addressed in Chapter 4:

→ The Roadmap begins with the **basic topics**, such as terminologies and definitions, classifications and ethical issues. They are the basis for AI discussions and are thus the central core of the Roadmap.

→ The **security/safety** of AI systems plays a crucial role in widespread use of AI solutions. Only a more in-depth consideration of requirements for operational safety and information security, for example, can enable the comprehensive use of AI systems in business and society.

→ Another key topic, and the basis for the broad market success of AI, is **testing and certification**. This requires reliable quality criteria and reproducible test methods that can be used to verify the properties of AI systems. They are a key prerequisite for assessing the quality of AI-based applications and contribute significantly to explainability and traceability – two factors that build trust and acceptance.

→ Another challenge in the use of AI, especially for small and medium-sized enterprises, is the integration of AI technologies in organizations. The focus here is on **socio-technical aspects** such as human-technology interaction, humane work design, and requirements for business structures and processes, which are all examined in the Roadmap.

→ The fields of application of AI are extremely diverse. AI technologies are used in almost all business and application areas and offer great potential. To cover a broad spectrum of applications, the Roadmap considers industry-specific challenges for the following five sectors in particular, in addition to the cross-cutting issues mentioned above: **Industrial Automation, Mobility, Medicine, Financial Services** and **Energy / Environment**.

The present Roadmap outlines the work and discussion results for all nine key topics and provides a comprehensive overview of the status quo, requirements, and needs for action.

With more than 116 identified standardization needs, the Roadmap illustrates concrete potential in all core topics and formulates six central recommendations for action in Chapter 2:

- → Develop, validate, and standardize a horizontal conformity assessment and certification programme for trustworthy AI systems
- → Establish data infrastructures and elaborate data quality standards for the development and validation of AI systems
- → Consider humans as part of the system in all phases of the AI life cycle
- → Develop specifications for the conformity assessment of continual or incremental learning systems in the field of medicine
- → Develop and deploy secure and trusted AI applications in mobility through best practices and assurance
- → Develop overarching data standards and dynamic modelling techniques for the efficient and sustainable design of AI systems

The high dynamics in AI technology development and the rapid increase in industrial applications of AI systems are also placing new types of demand on standardization processes and on the provision and further utilization of standards content. To meet these challenges, the standards organizations are developing new approaches, which are listed in Chapter 5 of the Roadmap. The focus here is on reviewing and adapting the current collections of standards, analyzing standardization needs, and the agile development and demand-driven provision of standards and specifications.

This Standardization Roadmap sets the path for future standardization in the field of artificial intelligence. Initial needs for action were already identified in the first edition of the Roadmap, a large number of which have been initiated or implemented as standardization and research projects. The current status of the implementation activities from the first edition is described in Chapter 6.

The publication of the second edition of the Standardization Roadmap represents the starting point for the implementation of the results. Here, too, it is important to launch standardization activities in line with the recommendations for action and to leverage the identified potential with the help of the resulting standards and specifications. Standards and specifications will support German industry and science in creating innovation-friendly conditions for the technology of the future. In particular, the results of the Roadmap can make an important contribution to the socio-political debate on the role and use of AI.

Standardization in Germany is based on the cooperation of experts from industry, science, the public sector and civil society. Only an early involvement of AI experts in standardization bodies will make it possible to incorporate German interests in International Standards and thus to develop market-oriented standards and specifications for AI, and also to strengthen Germany's position as an economic nation and exporting country.

1 Introduction

From robots working together in Industrie 4.0 to intelligent voice assistants and autonomous driving cars – artificial intelligence is changing our economy and society in a lasting manner. The self-learning and continuously improving AI systems enable more efficient processes in production and other areas. Completely new business models can be created through such systems. The possibilities are endless – and yet such an influential technology should stay within certain boundaries in order to actually help us. Reliable, functional and, above all, safe and secure AI needs certain rules: first of all, a common understanding and a uniform language, so that everyone is talking about the same thing. In addition, open interfaces are necessary for the systems to realize their full potential and work together efficiently. Only in this way can different AI-controlled machines communicate with each other, and products become visible along the entire value chain. At the same time, ethical issues play a central role in the use of artificial intelligence. Distortion, discrimination, and manipulation should be prevented from the outset if AI is to benefit humans.

Standards and specifications make a key contribution to all these aspects: They define requirements for artificial intelligence and structure the technology landscape. This makes them a strategically important instrument for strengthening the innovation and competitiveness of the German economy. The estimated economic benefit of standards is around 17 billion euros a year [1]. This is one of the reasons why work on the second edition of the Standardization Roadmap Artificial Intelligence was begun. The task of this Roadmap is to formulate a strategic plan for AI standardization.

1.1 Role of standardization in the field of AI

The development of AI applications has advanced rapidly in recent years, and so creating a future-proof framework for AI is essential.

Standards and specifications play an important part in this. They enable a reliable and safe application of AI technologies and contribute to explainability and traceability. This in turn makes them key factors for the acceptance of AI applications and creates trust in the market and among consumers.

Standards and specifications are a crucial factor for the broad market success of AI: They help to establish innovations more quickly by promoting the rapid transfer of technologies from research to application, thus making it easier for German companies to enter European and international markets. Small and medium-sized companies in particular benefit from this, because open interfaces and uniform requirements make it easier for them to access international markets.

Those who get involved in the development of standards and specifications can actively shape the global technical rules for AI and thus gain a lead. Early engagement of German stakeholders in national, European and international standardization is therefore essential to strengthen Germany as a global economy and exporting nation. International competitors have recognized this advantage; China and the USA in particular are major drivers of international AI standardization. If Germany and its European partners want to ensure that European values and ethical guidelines are adequately reflected in international AI standards, participation in standardization and an increased presence in international AI standardization bodies is strongly advised.

Policymakers have also recognized standardization as a strategic instrument for international competitiveness. This is one of the reasons why the German government has identified standardization as a central element in its AI Strategy (see Chapter 1.3). The European Commission also published a proposal for a legal framework for AI in spring 2021: the Artificial Intelligence Act (AI Act) (see Chapter 1.4). With the world’s first legal framework for AI, the EU aims to ensure the safety and fundamental rights of people and businesses when using AI, while strengthening investment and innovation. The planned AI Act assigns a central role to standardization: Harmonized European Standards are to technically specify requirements for transparency, robustness, and accuracy in the future, particularly in the area of high-risk AI applications.

1.2 Objectives and content of the Standardization Roadmap AI

Standards and specifications make a central contribution when it comes to defining requirements for artificial intelligence and structuring the technology landscape.

The early development of a strategic plan that identifies standardization needs and makes recommendations is essential. The Standardization Roadmap AI represents such a strategic plan. It describes a framework for action for standardization in the field of AI, which is created on the basis of a broad coordination process and is thus the essential foundation for initiating corresponding work in standardization at national, but above all at European and international level.

The Standardization Roadmap AI thus implements a key **measure of the German government's AI Strategy**] [2] and makes a significant contribution to introducing the national position at an early stage at European and international level, thereby decisively strengthening Germany's role as an economic nation and exporting country. The aim of the Standardization Roadmap AI is to create innovation-friendly framework conditions for the technology of the future and to support German industry and science in the international competition for the best solutions and products in the field of AI.

The Standardization Roadmap is to be understood as a "living document" that presents the current results of work and discussions, and serves as a central communication medium for exchange between standardization bodies, industry, associations, research institutions, civil society and policymakers.

It is developed and regularly updated in an open, transparent and broad-based participation process by representatives from business, science, the public sector and civil society.

1.2.1 Objectives of the Standardization Roadmap AI

The Roadmap has two **main objectives**: First, it describes the environment in which AI standardization is taking place and provides an overview of existing standards and specifications on aspects of AI. Second, it identifies standardization needs and formulates concrete recommendations for action. The Roadmap thus sets the path for future standardization in the field of AI and makes a significant contribution to establishing "AI – Made in Germany" as a strong brand for developing new business models, effective innovations and scalable applications. This makes it the signpost for AI standardization and at the same time offers great potential for raising European values to the international level.

The target audience of the Standardization Roadmap is the broad AI professional public. Its recommendations are primarily directed at industry, but also at representatives of the quality infrastructure, politics, research and civil society.

First edition of the Standardization Roadmap AI

DIN and the German Commission for Electrical, Electronic & Information Technologies in DIN and VDE (DKE), on behalf of the Federal Ministry for Economic Affairs and Climate Action (BMWK), had initiated work on the first edition of the German Standardization Roadmap "Artificial Intelligence" as early as the end of 2019. With the participation of more than 300 experts from various fields, the first edition of the Standardization Roadmap AI was developed and presented to the expert public for the first time at the German government's Digital Summit in November 2020. The results of the Standardization Roadmap AI represent an inventory and serve as a strategic plan for standardization in the field of AI. Since its publication, work has been proceeding at full speed to implement the total of 78 identified recommendations for action and needs from the seven main topics (basic topics, ethics, quality/certification/conformity assessment, IT security, industrial automation, mobility/logistics and medicine). A large number of standardization projects have been successfully initiated and also introduced at European and international level (see Chapter 6).

In addition, "Lighthouse Projects of the German Standardization Roadmap AI" have been launched to implement the overarching recommendations for action of the Roadmap (1st edition). With the help of these implementation projects, practical experience is being gathered in the context of the respective application, concrete standardization needs are being derived, and findings on quality and conformity testing are being gained. The "lighthouse projects" are thus of particular importance in the implementation of the Standardization Roadmap AI, which is why they enjoy increased attention among standardization stakeholders and are widely visible in industry, research and politics. An overview of the standardization projects initiated and the status of the implementation activities of edition 1 of the Roadmap is presented in Chapter 6.

Second edition of the Standardization Roadmap AI

The high dynamics in AI research and industrial development and application on the one hand, and emergent changes at the regulatory level on the other, require continuous further development of the strategic framework for action and the recommendations of the Standardization Roadmap AI. For this reason, work on the second edition of the Standardization Roadmap Artificial Intelligence was launched in January 2022 on behalf of the Federal Ministry for Economic Affairs and Climate Action (BMWK). The aim is to continue and further develop the previous results. Thus, this Roadmap builds on the findings of the first edition and is considered a standalone document.

The Standardization Roadmap AI focuses on selected key topics: Horizontal and cross-cutting aspects as well as sector-specific challenges are highlighted. In addition to the previous topics such as basic topics, security/safety, testing and certification, industrial automation, mobility as well as medicine, the focus will additionally be on the new aspects sociotechnical systems, financial services and energy/environment. The task of the Roadmap is to provide a comprehensive overview of the status quo, requirements and challenges, as well as standardization needs for the nine key topics mentioned above. Within the scope of the second edition of the Standardization Roadmap AI, industry-relevant use cases are considered in detail and concrete standardization needs are derived from them.

In addition, this Standardization Roadmap AI pays special attention to the draft Artificial Intelligence Act (AI Act) published by the European Commission. This planned, world's first legal framework for AI assigns a central role to standardization. In the area of high-risk AI applications in particular, requirements for AI systems are to be technically specified in future in harmonized European Standards (see Chapter 1.4). One task of the second edition of the Standardization Roadmap AI is therefore also to identify needs for standards and specifications to implement the AI Act and to take these into account when designing the further ground plan for standardization.

The publication of the Standardization Roadmap AI is being followed by the phase of implementation and stabilization of the results: Within the framework of implementation and "lighthouse" projects, practical experience is to be gathered for typical and industry-relevant AI applications, requirements are to be identified, and concrete standardization needs are to be derived and implemented. The central objective of consolidation is to incorporate the identified topics and needs (see Chapter 4) in the relevant standardization bodies, to initiate concrete standardization activities and ultimately to develop standards and specifications. The focus for 2023 and 2024 will therefore be on consolidating the results of the Standardization Roadmap.

It should be noted that standardization is always a joint task based on the broad participation and collaboration of experts from industry, science, government and society as a whole. Only the early engagement of experts with broad experience and insights from practice will make it possible to develop standards and specifications for AI that are in line with the market and with needs, and to ensure their acceptance in industry, science and society as a whole. If Germany wants to ensure that its interests are adequately reflected in international AI standards, active participation in standardization and increased presence in international AI standardization bodies are strongly advised.

1.2.2 Coordination Group AI Standardization and Conformity

The work on the Standardization Roadmap AI and its implementation is steered and accompanied by the high-level Coordination Group "AI Standardization and Conformity"[1] (referred to below as the Coordination Group). This was founded in May 2021 with a mandate from the German government, represented by the Federal Ministry for Economic Affairs and Climate Action (BMWK), the Federal Ministry of Education and Research (BMBF) and the Federal Ministry of Labour and Social Affairs (BMAS), and is composed of leaders from all relevant areas for AI. The 17 members from industry, standardization, politics, science and civil society represent important topics, disciplines, industries and companies of different sizes, and see themselves as ambassadors for AI standardization (see Figure 1). The Coordination Group thus replaces the "Steering Group Standardization Roadmap AI", which accompanied and steered the previous activities for the first edition of the Standardization Roadmap.

[1] www.din.de/go/koordinierungsgruppe-ki

The Coordination Group is responsible for the content and strategic direction of the Roadmap, provides impetus for important innovation and socio-political developments, and advocates for national and international cooperation in the field of AI. It also drives forward the practical implementation of the Roadmap's recommendations in a targeted manner and coordinates all activities arising from them. At the same time, it serves as a general contact point for standardization on the topic of artificial intelligence and as a place where the entire German AI landscape can coordinate, exchange and participate. The Coordination Group is supported by the Expert Group, whose 24 members[2] have been appointed to support the Coordination Group members and/or to network with relevant initiatives or actors.

1.2.3 Methodical approach

The work on the second edition of the Standardization Roadmap AI kicked off with a virtual event[3] on 20 January 2022, attended by more than 600 participants (see Figure 2). Speakers from politics, business, standardization, science and civil society organizations discussed the objectives and approach of the Roadmap and provided thematic insights into the key topics. In addition, practical implementation projects that emerged from the recommendations for action in the first edition of the Roadmap AI were highlighted.

As with the first edition of the Roadmap, the participation of experts from all stakeholders is the essential basis for the development of the Standardization Roadmap.

Interested representatives from industry, science, the public sector and civil society, as well as representatives of groups already constituted and involved in the topic of AI, were invited to actively contribute their expertise to the development of the Standardization Roadmap. In this context, the consideration of different perspectives and associated requirements is of great importance, so that both technical and non-technical aspects were equally incorporated into the development process of the Standardization Roadmap AI.

More than 570 experts from various industries and with different backgrounds of experience were recruited to participate and contributed their expertise. The Roadmap was developed in nine working groups on various key topics (see Chapter 4) and was organized completely virtually on the DIN.ONE[4] collaboration platform. Figure 3 shows the composition of the working groups.

[2] The members of the Expert Group are: Dir. u. Prof. Dr. Lars Adolph (BAuA), Nikolas Becker (GI), Dr. Tarek R. Besold (DEKRA DIGITAL), Jens Brinckmann (BMWK), Egbert Fritzsche (VDA), Dr. Patrick Gilroy (TÜV-Verband), Dr. Sebastian Hallensleben (VDE), Taras Holoyad (Bundesnetzagentur), Dr. Maximilian Hösl (acatech), Dr. Jürgen Klippert (IG Metall), Alena Kühlein (DIHK), Daniel Loevenich (BSI), Dr. Christoph March (BMBF), Manfred Meiss (BMWK), Dr. Maximilian Poretschkin (Fraunhofer IAIS), Prof. Dr. Georg Rehm (DFKI), Guido Reimann (VDMA), Jochen Reinschmidt (ZVEI), Dr. Kinga Schumacher (DFKI), Rosmarie Steininger (Chemistree), Dr. Christina Strobel (KI Bundesverband), Hauke Timmermann (eco – Verband der Internetwirtschaft), Merle Uhl (Bitkom e. V.), PD Dr. Marc Wittlich (IFA der DGUV)

[3] Exerpts from the event are on YouTube: https://youtu.be/jt_xen012xU; for more information on the event go to: www.din.de/go/auftakt-ki)

[4] www.din.one/site/ki

MEMBERS OF THE COORDINATION GROUP:

Dr. Daniela Brönstrup
Federal Ministry for Economic Affairs and Climate Action (BMWK)

Dr. Joachim Bühler
TÜV Association

Dr. Detlef Gerst
IG Metall

Dr. Tobias Heimann
ZVEI Germany's Electro and Digital Industry and Siemens Healthineers

Dr. Wolfgang Hildesheim
Bitkom and
IBM Deutschland

Dr. Vanessa Just
German AI Association

Julia Kloiber
Superrr Lab

Prof. Antonio Krüger
German Research Centre for Artificial Intelligence (DFKI)

Dr. Christoph Peylo
Machinery and Equipment Manufacturers Association (VDMA) / German Association of the Automotive Industry (VDA) and Robert Bosch GmbH

Alexander Rabe
eco Association of the Internet Industry

Prof. Ina Schieferdecker
Federal Ministry of Education and Research (BMBF)

Dr. Volker Treier
Association of German Chambers of Industry and Commerce (DIHK)

Prof. Wolfgang Wahlster
Germany's Platform For Artificial Intelligence and German Research Centre for Artificial Intelligence (DFKI)

Prof. Dieter Wegener
German Commission for Electrical, Electronic & Information Technologies of DIN and VDE (DKE)

Christoph Winterhalter
German Institute for Standardization (DIN)

Prof. Stefan Wrobel
Fraunhofer Institute for Intelligent Analysis and Information Systems (IAIS)

PERMANENT GUESTS OF THE COORDINATION GROUP:

Dr. Johannes Winter[5]
Plattform Lernende Systeme – Germany's Platform For Artificial Intelligence

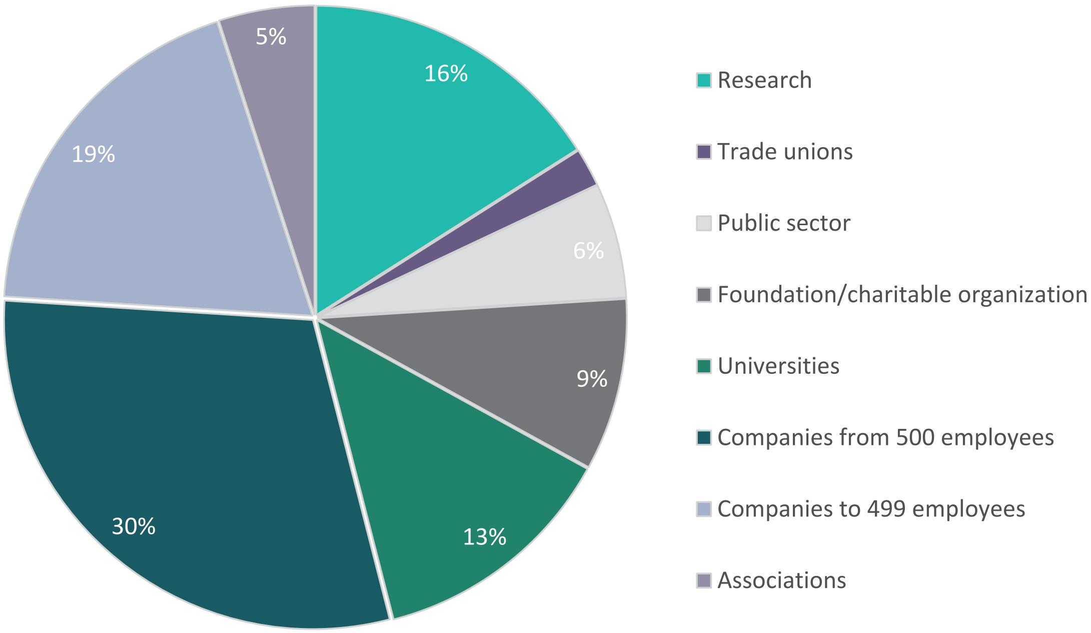
Figure 1: Members of the Coordination Group AI Standardization and Conformity (Source: DIN)

[5] to 08/2022 "Platform Learning Systems", from 09/2022 L3S AI Research Centre

DIN DKE
HERZLICH WILLKOMMEN ZUR
AUFTAKTVERANSTALTUNG
Normungsroadmap Künstliche Intelligenz – Ausgabe 2
Die Veranstaltung beginnt um 10:00 Uhr
>> Drücken Sie bitte „Play" <<
Normungsroadmap KI - Der Initialzünder für neue Projekte
safe.trAIn – Sichere KI am Beispiel fahrerloser Regionalzug
Dr. Cornel Klein, Siemens AG
Dr. Daniela Brönstrup
BMWK
Dr. Anna Christmann
Mitglied des Deutschen Bundestags
Auftaktveranstaltung Normungsroadmap Künstliche Intelligenz – Ausgabe 2
ZERTIFIZIERTE KI
Dr. Maximilian Poretschkin
Fraunhofer IAIS

Figure 2: Scenes from the kick-off event (Source: DIN)

Figure 3: Composition of the nine working groups of the Standardization Roadmap AI (Source: DIN)



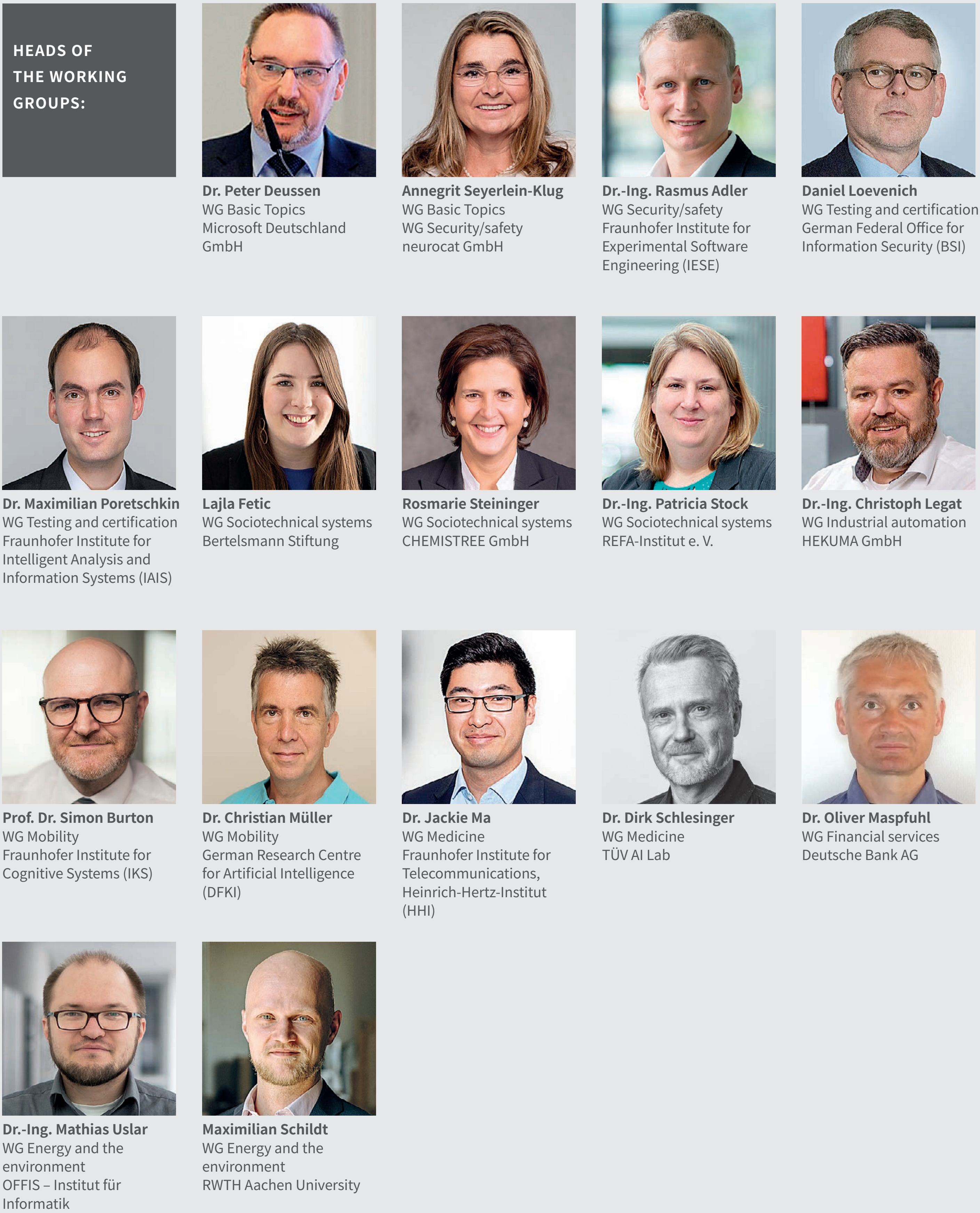
Experienced experts were recruited to chair the working groups (see Figure 4), who led the substantive work and reported regularly to the Coordination Group and the group of experts.

1. Basic topics (Heads: Dr. Peter Deussen, Microsoft Deutschland GmbH, and Annegrit Seyerlein-Klug, neurocat GmbH)
2. Security/safety (Heads: Dr.-Ing. Rasmus Adler, Fraunhofer-Institut für Experimentelles Software Engineering (IESE), and Annegrit Seyerlein-Klug, neurocat GmbH)
3. Testing and certification (Heads: Dr. Maximilian Poretschkin, Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme (IAIS), and Daniel Loevenich, Federal Office for Information Security (BSI)
4. Sociotechnical systems (Heads: Rosmarie Steininger, CHEMISTREE GmbH, Dr.-Ing. Patricia Stock, REFA-Institut e. V., and Lajla Fetic, Bertelsmann Stiftung)
5. Industrial automation (Head: Dr.-Ing. Christoph Legat, HEKUMA GmbH)
6. Mobility (Heads: Prof. Dr. Simon Burton, Fraunhofer-Institut für Kognitive Systeme (IKS), and Dr. Christian Müller, Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI))
7. Medicine (Heads: Dr. Jackie Ma, Fraunhofer-Institut für Nachrichtentechnik, Heinrich-Hertz-Institut (HHI), and Dr. Dirk Schlesinger, TÜV AI Lab)
8. Financial services (Head: Dr. Oliver Maspfuhl, Deutsche Bank AG)
9. Energy/environment (Heads: Dr.-Ing. Mathias Uslar, OFFIS – Institut für Informatik, and Maximilian Schildt, RWTH Aachen University)

Figure 5 shows the overall structure of the project for the Standardization Roadmap AI.

This Standardization Roadmap AI was published at the end of 2022 and handed over to the German government. It is available to be downloaded for free in German and in English at https://www.din.de/go/roadmap-ai.

The publication of the Standardization Roadmap AI marks the immediate start of the implementation and consolidation of the results. Then, with the support of all Federal Ministries and the participation of experts from industry, research and civil society, as many of the recommendations for action as possible must be implemented quickly in the form of concrete implementation projects and standardization activities.



HEADS OF THE WORKING GROUPS:

Dr. Peter Deussen
WG Basic Topics
Microsoft Deutschland GmbH

Annegrit Seyerlein-Klug
WG Basic Topics
WG Security/safety
neurocat GmbH

Dr.-Ing. Rasmus Adler
WG Security/safety
Fraunhofer Institute for Experimental Software Engineering (IESE)

Daniel Loevenich
WG Testing and certification
German Federal Office for Information Security (BSI)

Dr. Maximilian Poretschkin
WG Testing and certification
Fraunhofer Institute for Intelligent Analysis and Information Systems (IAIS)

Lajla Fetic
WG Sociotechnical systems
Bertelsmann Stiftung

Rosmarie Steininger
WG Sociotechnical systems
CHEMISTREE GmbH

Dr.-Ing. Patricia Stock
WG Sociotechnical systems
REFA-Institut e. V.

Dr.-Ing. Christoph Legat
WG Industrial automation
HEKUMA GmbH

Prof. Dr. Simon Burton
WG Mobility
Fraunhofer Institute for Cognitive Systems (IKS)

Dr. Christian Müller
WG Mobility
German Research Centre for Artificial Intelligence (DFKI)

Dr. Jackie Ma
WG Medicine
Fraunhofer Institute for Telecommunications, Heinrich-Hertz-Institut (HHI)

Dr. Dirk Schlesinger
WG Medicine
TÜV AI Lab

Dr. Oliver Maspfuhl
WG Financial services
Deutsche Bank AG

Dr.-Ing. Mathias Uslar
WG Energy and the environment
OFFIS – Institut für Informatik

Maximilian Schildt
WG Energy and the environment
RWTH Aachen University

Figure 4: Heads of the working groups (Source: DIN)

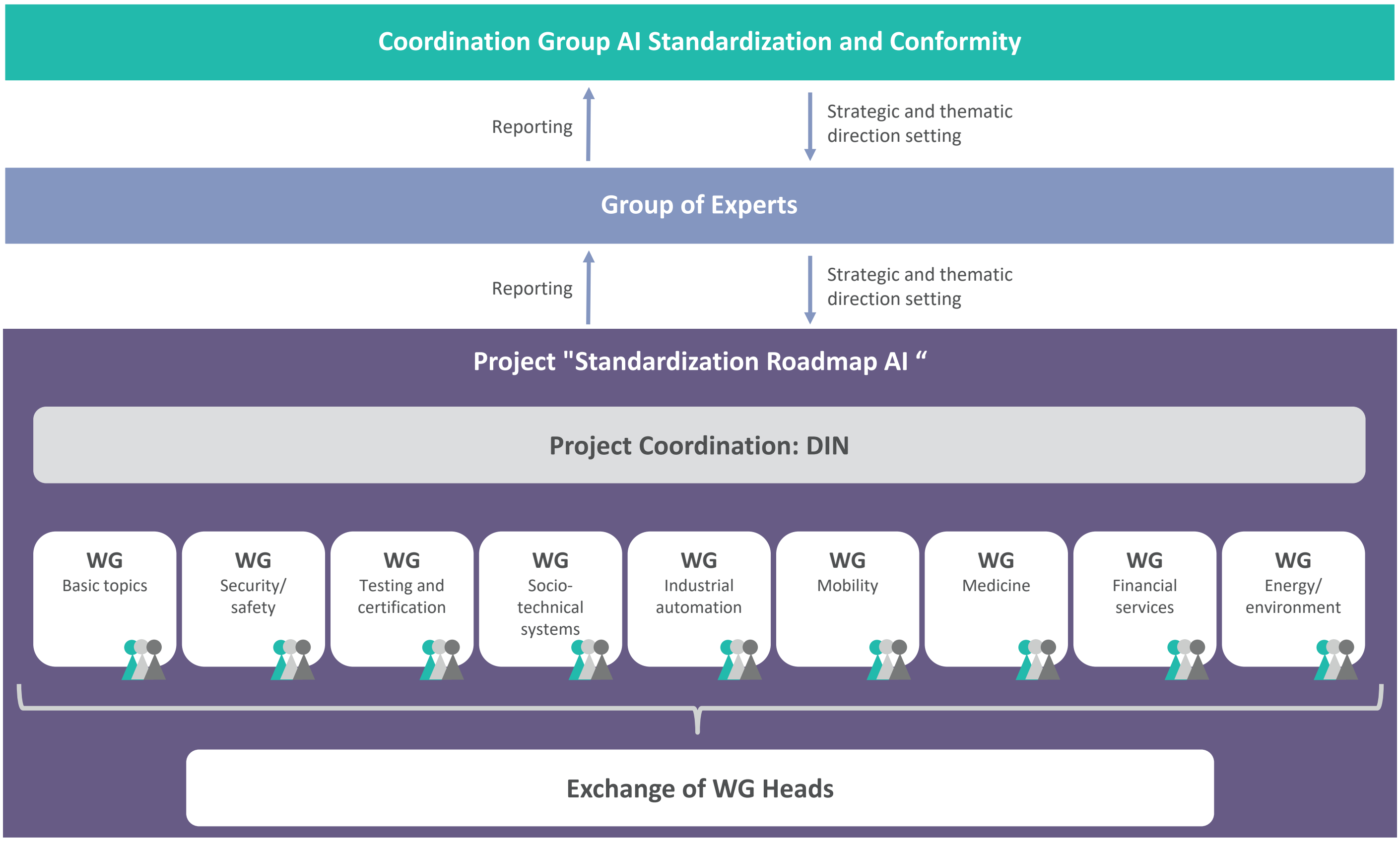

Coordination Group AI Standardization and Conformity
Reporting
Strategic and thematic direction setting
Group of Experts
Reporting
Strategic and thematic direction setting
Project "Standardization Roadmap AI "
Project Coordination: DIN
WG Basic topics
WG Security/ safety
WG Testing and certification
WG Socio-technical systems
WG Industrial automation
WG Mobility
WG Medicine
WG Financial services
WG Energy/ environment
Exchange of WG Heads

Figure 5: Project structure of the Standardization Roadmap AI (Source: DIN)

1.3 AI Strategy of the German Federal Government

In November 2018, the German government adopted its national strategy “Artificial Intelligence” [2] and intends to use it to further develop Germany into a leading location for AI and strengthen the competitiveness of the German and European economies, especially vis-à-vis the USA and China. In this context, the potential of human-centric AI is to be exploited in accordance with the European economic, value and social structure in order to promote the application of AI on a broad scale.

By updating the national AI Strategy at the end of 2020, the German government responded to new developments and needs that have arisen since the first edition was published [3]. The update focuses on developments in the wake of the Covid 19 pandemic, sustainability issues, in particular environmental and climate protection, and European and international networking. The financial resources for implementing the strategy by 2025 have been increased from the previous three billion euros to five billion euros.

Specifically, with the update of the AI Strategy, the German government intends to

→ train, recruit and retain more AI specialists in Germany,

→ establish high-performance and internationally visible research structures and, in particular, provide internationally competitive cutting-edge AI and computing infrastructures,

→ establish AI ecosystems with an international reach based on excellent research and transfer structures in order to accelerate the application of research results in business practice, especially in the SME sector, and to boost the start-up dynamics,

→ strengthen the underlying conditions for innovative and human-centric AI applications in Germany and Europe by establishing and expanding the quality infrastructure on the basis of an appropriate regulatory framework into a system for safe, secure and trustworthy AI, and

→ support civil society networking and involvement in the development and use of AI that serves the common good [3].

Even in the first edition of the AI Strategy, the German government assigns a central role to standards and presents standardization as a central building block of the Strategy. It states that the German government will (among other things) develop a Roadmap on standards and specifications in the field of AI in a joint project with DIN. Furthermore, the review of existing standards and specifications for "AI suitability" as well as the development of machine-readable and machine-interpretable standards and specifications (SMART standards) for AI applications is suggested.

The update [3] also clearly emphasizes and elaborates the importance of standards and specifications in the field of AI. It states:

→ "By setting clear rules and standards, the fundamental rights of citizens can be protected, trust in AI can be strengthened, sustainable deployment of AI as well as innovation and competition can be promoted." (p.6)

→ "This forms the basis for a subsequent implementation programme, which, building on the Roadmap, is to initiate specific standardization projects, address certification issues for learning systems and initiate the rapid transferability of the findings gained into international standards and test criteria. Key topics here are inter alia safety and security, robustness, transparency and non-discrimination in AI systems" (p.19)

→ "In combination with metrology, accreditation, conformity assessment, market surveillance and environmental audits, rules, norms and standards form the quality infrastructure – the backbone of the 'Made in Germany' brand." (p. 20)

→ "Implementing the roadmap defined in the Standardization Roadmap AI: developing test criteria on the basis of established and future test technologies to test the robustness, safety and security, reliability, integrity, transparency, explainability, interpretability and non-discrimination of (hybrid) AI systems". (p. 31)

By highlighting and promoting the development of technical standards and specifications within the framework of the national AI Strategy, economic processes will be facilitated, technology transfer will be favoured, and confidence in AI products and AI services will be strengthened via the national quality infrastructure.

1.4 AI regulation at European level

In April 2021, the European Commission published a landmark draft to regulate the use of artificial intelligence – the Artificial Intelligence Act (AI Act) [4]. The draft regulation represents the world's first regulatory framework for AI and is based on the "Coordinated Plan on Artificial Intelligence" [5], the "Policy and Investment Recommendations for Trustworthy Artificial Intelligence" [6] and the "White Paper on Artificial Intelligence" [7]. The stated goal of the planned AI Act is to ensure fundamental rights and security in the European Union when using AI, while promoting investment and innovation in the EU member states.

In the draft, the Commission basically assumes a very broad concept of artificial intelligence and pursues a technology-neutral and risk-based approach. Based on the recommendations of the independent High Level Expert Group (HLEG) set up by the Commission, three essential components for trustworthy artificial intelligence have been defined: lawfullness, ethics and robustness. Four ethical principles were identified as the foundations of trustworthy AI: Respect for human autonomy, prevention of harm, fairness, and explainability. The realization of these principles has been described in seven core requirements (see also Chapter 4.1.2.1):

→ Huan agency and oversight
→ Technical robustness and safety
→ Privacy and data governance
→ Transparency
→ Diversity, non-discrimination and fairness
→ Societal and environmental well-being
→ Accountability [8]

The planned AI Act supports the recommendations of the HLEG. Globally, the Act aims to provide European leadership in the development of safe, trustworthy and ethical AI. European Standards will play an important role in the technical design of the legal requirements of the Act.

1.4.1 Scope

The legislative proposal is currently being discussed in the European Council and the European Parliament. The trialogue between the three legislative institutions (European Commission, European Council and European Parliament) is preparing for the adoption of the Act. The AI Act is expected to be enacted in 2022/2023 and will take effect 24 months later in 2024/2025 [9].

Table 1: Scope of and penalties under the proposed AI Act (Status: Commission draft [4])

Scope	a) Providers placing on the market or putting into service AI systems in the Union, irrespective of whether those providers are established within the Union or in a third country b) Users of AI systems located within the Union c) Providers and users of AI systems that are located in a third country, where the output produced by the system is used in the Union
Regulatory focus	High-risk AI systems
Reference to standards	The draft regulation includes reference to harmonized European Standards developed by the European standards organizations on the basis of a standardization mandate from the European Commission for the technical design of the essential requirements for high-risk AI systems (Art. 40 of the Commission draft).
Penalties	a) If the offender is a company, 6 % of its total worldwide annual turnover for the preceding financial year or 30 million euros – whichever is higher (for non-compliance with Art. 5 and Art. 10) b) Up to 4 % of its total worldwide annual turnover for the preceding financial year or 20 million euros – whichever is higher (for non-compliance)
Relevant AI systems	All AI systems (specified in Annex 1)
Timeline	→ Expected final deliberation in the European Parliament: Q4 2022 at the earliest → Expected final deliberation in the European Council: Q4 2022 at the earliest → Expected trialogue: 2023 → Entry into force (according to current draft): 20 days after adoption → Implementation (according to current draft): 24 months after entry into force

The "Artificial Intelligence Act" is to be enacted as a regulation. A regulation is a binding legal act of the European Union with general validity and direct effect in all member states; transposition into national law is not required. Civil law issues in the use of AI (e.g., liability, attribution of declarations of intent, creation of intellectual property, etc.) are not addressed by the draft regulation. It is primarily an act that prohibits the use of AI systems in certain application scenarios or makes it dependent on technical and organizational requirements. The technical requirements for permitted high-risk AI systems are to be specified in harmonized European Standards. Art. 3 (1) of the draft Regulation defines an AI system as "software that is developed with one or more of the techniques and approaches listed in Annex I and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with". This definition can still be changed in the trilogue with the European Parliament and the European Council.

Annex I further defines these techniques and concepts identified in the proposed definition as follows:[6]

- → Machine learning approaches, including supervised, unsupervised and reinforcement learning, using a wide variety of methods including deep learning
- → Logic- and knowledge-based approaches, including knowledge representation, inductive (logic) programming, knowledge bases, inference and deductive engines, (symbolic) reasoning and expert systems
- → Statistical approaches, Bayesian estimation, search and optimization methods

According to Art. 4 and 73 of the draft Regulation, the European Commission is empowered to adapt Annex I at any time to market and technical developments.

[6] See AI Act draft, Annex I [4].

The planned AI Act is therefore a forward-looking reaction of the European Union to the increasing number of production and services that work with artificial intelligence technologies and that will be placed on the European market in the future. The authors of the Standardization Roadmap welcome this intention.

1.4.2 Legislative environment

The planned AI Act is integrated into a number of other laws at the EU level that must also be observed in the development of AI systems. These address issues in the areas of access to and use of data and related service structures (e.g., the planned Data Act, Data Governance Act, Digital Services Act, Digital Markets Act), data security issues (e.g., the General Data Protection Regulation, Cybersecurity Act, or the planned Cyber Resilience Act), and more general regulations such as those relating to product liability or occupational health and product safety (e.g., Product Liability Directive, Framework Directive – Occupational Safety and Health, or Machinery Directive). In addition, there are sectoral components of legislation, such as the Medical Devices Regulation in relation to the safety of products or the European Health Data Space in relation to access to data in the respective area of application. In addition, the EU Charter of Fundamental Rights should be mentioned as a central building block for the European legal order. A large number of the above-mentioned laws are already in force, while others are currently in the legislative process.

Figure 6 provides an overview of the interrelationships between the proposed AI Act and EU laws related to it. The focus of the presentation is the planned AI Act with the areas "General & Product Safety", "Data Protection (Security and Privacy)" and "Data and Services". The numbering used refers to the more detailed description in Table 16 (see Chapter 13.1). Regulations and standards that go beyond EU laws are indicated in the overview. Particular importance is attached to the harmonized European Standards, which are being developed by the European Committee for Standardization on the basis of a mandate from the EU Commission to specify technical requirements of the legal act and which represent a central reference for the implementation of the requirements of the planned AI Act.

Existing European standards that are not "harmonized" in the legal sense are often derived from International Standards, e.g., the IEEE P7000TM series ([10], [11], [12], [13]) or the standards of subcommittee ISO/IEC JTC1/SC 42 [14], such as the Technical Report ISO/IEC TR 24368:2022 [15], which provides an overview of ethical or social concerns in the use of AI components.

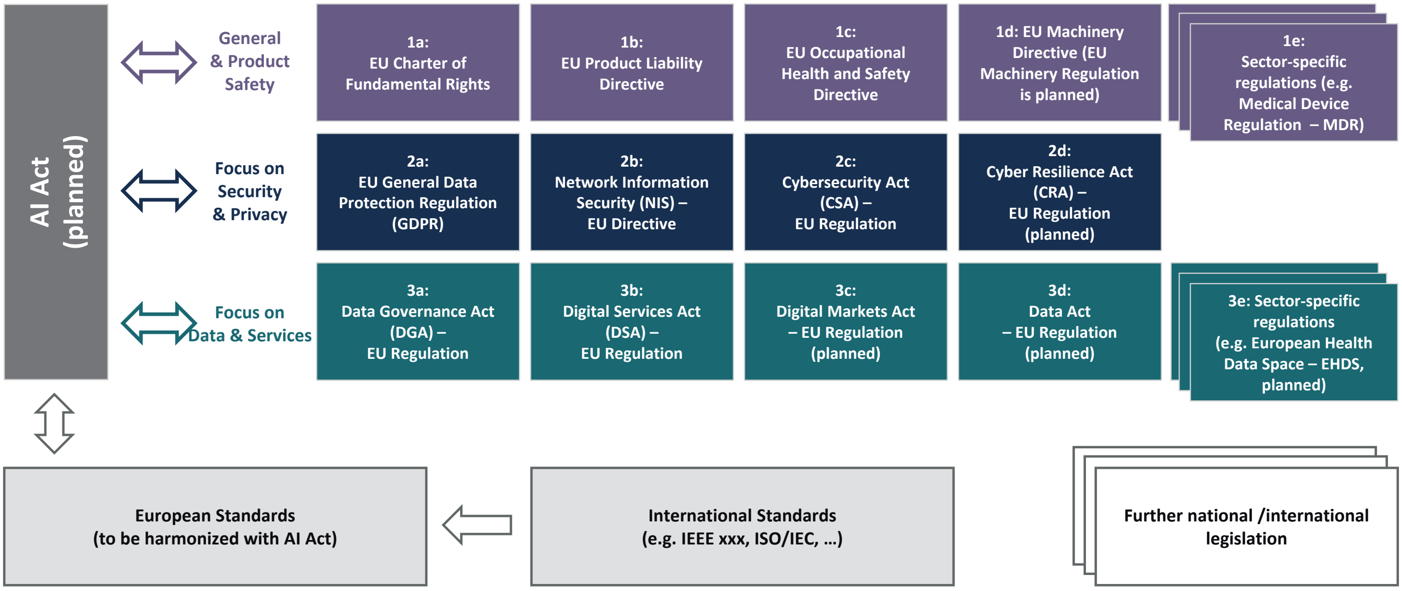

AI Act (planned)
General & Product Safety
Focus on Security & Privacy
Focus on Data & Services
1a: EU Charter of Fundamental Rights
1b: EU Product Liability Directive
1c: EU Occupational Health and Safety Directive
1d: EU Machinery Directive (EU Machinery Regulation is planned)
1e: Sector-specific regulations (e.g. Medical Device Regulation – MDR)
2a: EU General Data Protection Regulation (GDPR)
2b: Network Information Security (NIS) – EU Directive
2c: Cybersecurity Act (CSA) – EU Regulation
2d: Cyber Resilience Act (CRA) – EU Regulation (planned)
3a: Data Governance Act (DGA) – EU Regulation
3b: Digital Services Act (DSA) – EU Regulation
3c: Digital Markets Act – EU Regulation (planned)
3d: Data Act – EU Regulation (planned)
3e: Sector-specific regulations (e.g. European Health Data Space – EHDS, planned)
European Standards (to be harmonized with AI Act)
International Standards (e.g. IEEE xxx, ISO/IEC, …)
Further national /international legislation

Figure 6: Overview of EU laws with special reference to the planned AI Act (Source: Martin Haimerl)[7]

[7] The list does not claim to be complete.

In the case of the other legal acts, a wide range of requirements arise for the implementation and operation of AI-based systems, which are explained in more detail in Table 16 (see Chapter 13.1). For example, the General Data Protection Regulation contains requirements regarding the collection and processing of personal data and related restrictions, or the right not to be prejudiced by a decision based solely on automated processing (Art. 22 GDPR) and thus to preserve human autonomy. Potential conflict situations also arise because the proposed AI Act requires unrestricted access to training, validation, and testing datasets under certain circumstances (Art. 64 of the AI Act). In particular, if personal content should still be present here, security gaps could arise.

However, the laws listed also contain support measures, such as in the Data Act regarding better utilization of data in general or in the European Health Data Space regarding a suitable infrastructure for access to medical data. These approaches are in line with the EU's efforts to promote the implementation of AI innovations through the AI Act. Overall, efforts are being made at the EU level to create a comprehensive framework for harmonization and thus also for legal certainty in dealing with AI-based systems.

1.4.3 Summary: Objectives of the planned AI Act

In summary, the following key objectives of the AI Act can be identified:

→ Anchoring European values in AI systems
→ Ensuring EU fundamental rights
→ Establishing national and supranational control bodies
→ Defining the ethical application of AI
→ Defining artificial intelligence and AI Systems
→ Introducing a uniform framework to prevent fragmentation
→ Setting conformity standards through mandatory CE marking
→ Ensuring legal certainty to promote innovation and investment in AI
→ Establishing safe and secure AI life cycles
→ Regulating high-risk AI
→ Setting up a central database for high-risk AI
→ Strengthening innovation in the field of artificial intelligence

The draft regulation also contains measures to promote innovation, such as the establishment of real laboratories and "sandboxes" (Title V), requirements for governance structures at the Union and national levels, such as the establishment of a European Committee for Artificial Intelligence, the creation of a Union-wide database for stand-alone high-risk AI systems, and the introduction of certain monitoring and reporting requirements for AI system providers (Titles VI, VII, and VIII). Title IX lays the groundwork for creating codes of conduct to incentivize providers of AI systems that are not high risk to voluntarily apply the mandatory requirements for high-risk AI systems.

1.4.4 Significance of harmonized European Standards for the implementation of the AI Act

Standards play an important role in the proposed AI Act. They serve to reliably implement the requirements of the AI Act and help make the development of AI systems more efficient and reliable. Particular importance is attached to the harmonized European Standards (hEN), especially those listed in the Official Journal of the European Union for the planned AI Act. If distributors of high-risk AI systems comply with these hEN, it is assumed that they thereby also comply with the corresponding requirements of the legal act covered by the respective standard. This "presumption of conformity" facilitates placement on the European internal market. The application of the standard remains voluntary, but the placing on the market of high-risk AI systems without the application of hEN is likely to involve increased technical documentation.

A harmonized European Standard is defined as a standard adopted by the European standardization organizations CEN, CENELEC and/or ETSI "on the basis of a request made by the Commission for the application of Union harmonization legislation".[8] The process of developing an hEN is shown in Figure 7.

In its annual work programme for European standardization, the European Commission announces the areas in which it intends to issue standardization requests to the European standards organizations in a given year. On this basis, the Commission prepares a draft standardization request, which is discussed with standards organizations, stakeholders and sector experts, and is consulted with Harmonized Standards Consultants, and adapted as necessary. The standardization

[8] See EU Regulation on Standardization (1025/2012) Art. 2 para. 1c.

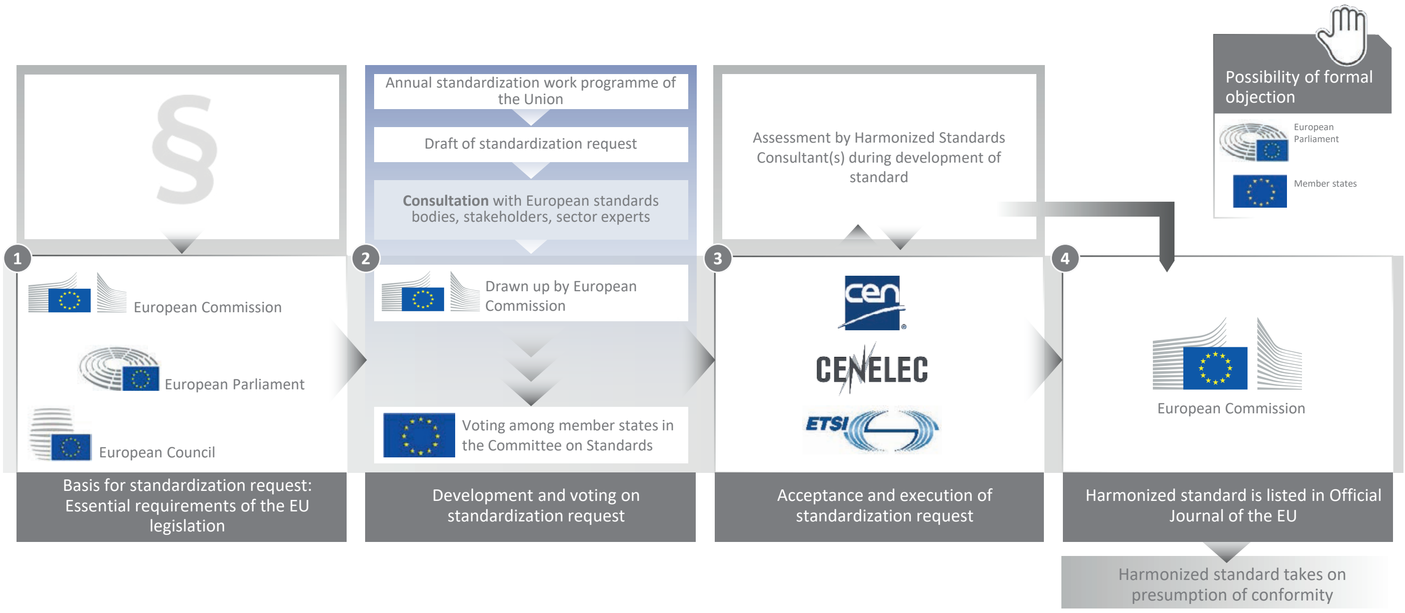

§
1
European Commission
European Parliament
European Council
Basis for standardization request: Essential requirements of the EU legislation
Annual standardization work programme of the Union
Draft of standardization request
Consultation with European standards bodies, stakeholders, sector experts
2
Drawn up by European Commission
Voting among member states in the Committee on Standards
Development and voting on standardization request
Assessment by Harmonized Standards Consultant(s) during development of standard
3
cen
CENELEC
ETSI
Acceptance and execution of standardization request
Possibility of formal objection
European Parliament
Member states
4
European Commission
Harmonized standard is listed in Official Journal of the EU
Harmonized standard takes on presumption of conformity

Figure 7: Process of developing harmonized European Standards (Source: DIN)

request is then discussed and voted on in the Committee on Standards, a body comprised of Member States[9]. Subsequently, the standardization request is handed over to the European standards organizations, which, after acceptance of the request, initiate the development of the standard within their bodies. The drafting process is accompanied by a Harmonized Standards (HAS) consultant financed by the EU Commission, who submits an assessment to the Commission as to whether the contents contained in the standard comply with the Commission's standardization request and technically reflect the essential requirements in the harmonization legislation. After completion of the standard, the EU Commission decides whether it is to be listed in the Official Journal of the European Union. Only with the listing of the standard in the Official Journal does it become a harmonized European standard and takes on the presumption of conformity. If a member state or the European Parliament considers that a listed hEN does not fully meet the requirements it is intended to cover, they may raise a formal objection to that hEN in accordance with Art. 11 of the Regulation on Standardization.

In May 2022, the EU Commission published a draft standardization request with which it intends to task the European standardization organizations with developing standards for the technical design of the essential requirements from Chapter 2 of the Commission's draft AI Act. The draft focuses in particular on the following topics as standardization requirements:

1. Risk management
2. Data and data governance
3. Record keeping through logging capabilities
4. Transparency and information for users
5. Human oversight
6. Accuracy specifications for AI systems
7. Robustness specifications for AI systems
8. Cybersecurity specifications for AI systems
9. Quality management system for providers of AI systems, including post-market monitoring process
10. Conformity assessment for AI systems

The planned adoption of the AI Act requires the timely development of harmonized European Standards. Since consensus-building in the standardization process takes time, it is important that the transition periods in the AI Act be sufficiently generous to ensure that all relevant standards are in place by the time the AI Act becomes mandatory. The planned AI Act indicates a second possibility. with Art. 41 (Common Specifications) of the Commission draft, which is intended to empower the European Commission to define technical specifications for the essential requirements itself by means of implementing acts. However, this must be viewed critically: It entails the risk of creating a parallel system to European standardization with competing technical requirements in terms of content, and does not demonstrate inclusivity and transparency in a comparable manner to the

[9] According to the EU Regulation on Standardization (1025/2012) Art. 22

European standardization system. Common Specifications according to Art. 41 should therefore only represent the last possible fallback option.

In this respect, the development of harmonized standards should continue to be the preferred way of technically shaping fundamental requirements of the planned AI Act in the future. International standards already in existence or under development (the IEEE 7000TM series [10], [11], [12], [13]) and ISO/IEC standards within the framework of JTC 1/SC 42 [14]) can be a good starting point for this. Further needs for action must be clarified in a targeted manner, particularly with regard to the requirements listed in the planned AI Act and specifically in the ten subject areas targeted in the draft standardization request. This Standardization Roadmap makes a central contribution to this.

1.4.5 Risk classification and structure of the AI Act

The proposed AI Act provides for the categorization of AI systems into four risk classes (see Figure 8):

→ AI systems that may not be placed on the market: Art. 5 para. 1 points a)–d)

→ High-risk AI systems: Art. 6 para. 1 in connection with Annex II, Clause A, no. 1–12, Art. 6 para. 1 in connection with Annex II, Clause B, no. 1–7 and Art. 6 para. 2 in connection with Annex III, no. 1–8

→ AI systems with special transparency obligations: Art. 52 para. 1–3

→ Low-risk AI systems: All AI systems that do not fall into any of the above groups

The primary classification in the AI Act is by industry or area of application (see Art. 5-7 and related Annexes). In the case of classification according to Art. 6, the classification is linked to the requirement that a third party conformity procedure is required in the sectoral harmonization regulations belonging to the respective sector and listed in Annex II (e.g. Machinery Directive, Medical Devices Regulation). In this respect, the classification in this case is indirectly co-determined by the sectoral harmonization regulation. Art. 7 and Annex III also list a number of use cases that are generally classified in the high-risk category.

The structure of the proposed AI Act is centrally oriented around these risk classes. Figure 9 provides an overview of the structure of the proposed AI Act with the Titles (I-XII) included and the relevant Articles (1-85). The diagram assigns the individual Titles to the respective risk classes. It indicates which requirements must be observed for which risk class. The requirements of the lower risk classes are transferred to the higher classes, i.e. they must also be applied there.

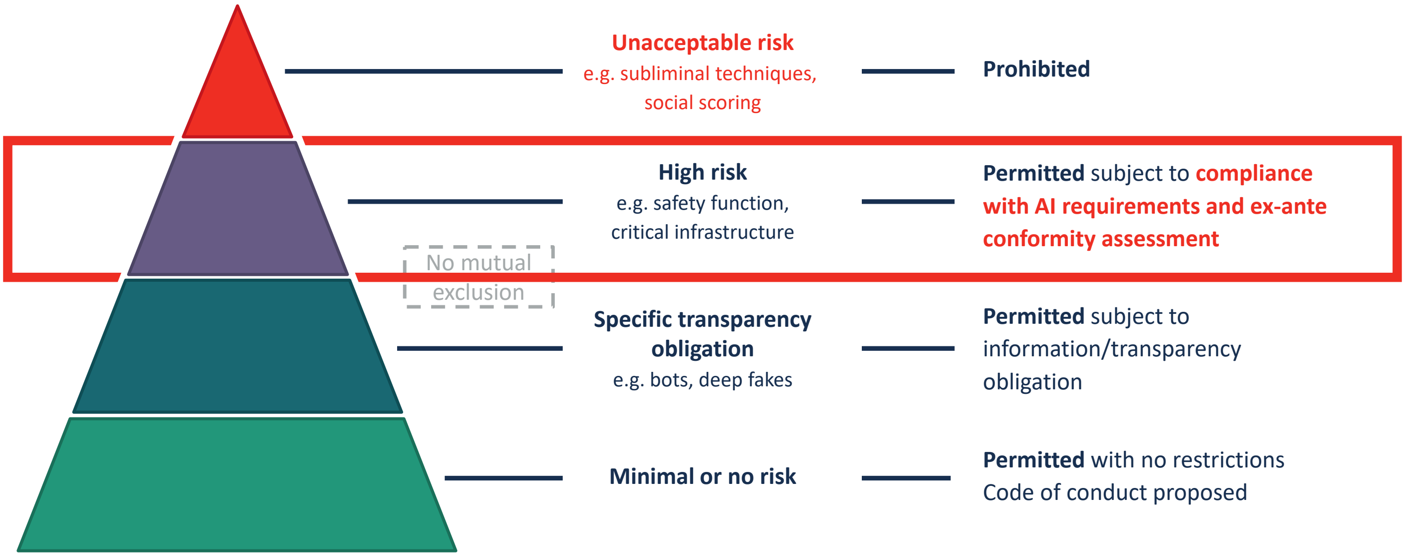
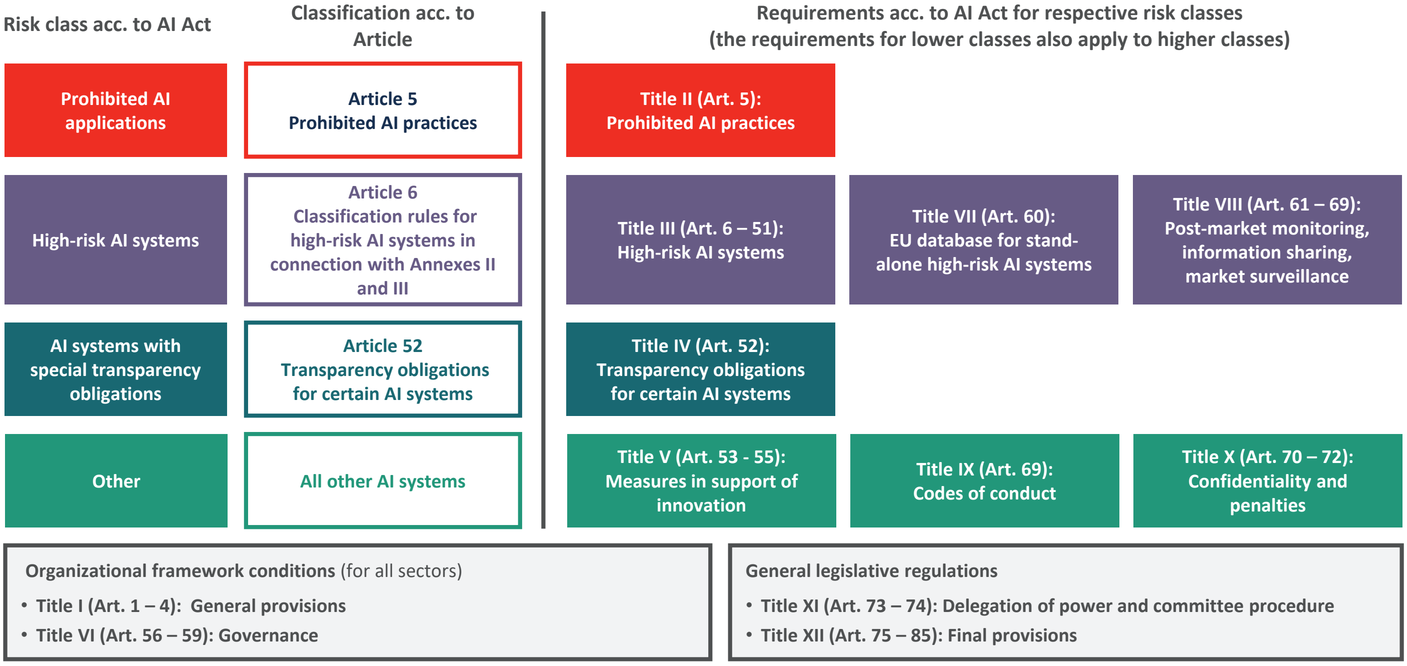

Unacceptable risk
e.g. subliminal techniques, social scoring
Prohibited
High risk
e.g. safety function, critical infrastructure
Permitted subject to compliance with AI requirements and ex-ante conformity assessment
No mutual exclusion
Specific transparency obligation
e.g. bots, deep fakes
Permitted subject to information/transparency obligation
Minimal or no risk
Permitted with no restrictions
Code of conduct proposed

Figure 8: Risk classes of the planned AI Act (Source: along the lines of [4])



Risk class acc. to AI Act	Classification acc. to Article	Requirements acc. to AI Act for respective risk classes (the requirements for lower classes also apply to higher classes)		
Prohibited AI applications	Article 5 Prohibited AI practices	Title II (Art. 5): Prohibited AI practices		
High-risk AI systems	Article 6 Classification rules for high-risk AI systems in connection with Annexes II and III	Title III (Art. 6 – 51): High-risk AI systems	Title VII (Art. 60): EU database for stand-alone high-risk AI systems	Title VIII (Art. 61 – 69): Post-market monitoring, information sharing, market surveillance
AI systems with special transparency obligations	Article 52 Transparency obligations for certain AI systems	Title IV (Art. 52): Transparency obligations for certain AI systems		
Other	All other AI systems	Title V (Art. 53 - 55): Measures in support of innovation	Title IX (Art. 69): Codes of conduct	Title X (Art. 70 – 72): Confidentiality and penalties

Organizational framework conditions (for all sectors)

- Title I (Art. 1 – 4): General provisions
- Title VI (Art. 56 – 59): Governance

General legislative regulations

- Title XI (Art. 73 – 74): Delegation of power and committee procedure
- Title XII (Art. 75 – 85): Final provisions

Figure 9: Overview of the content of the planned AI Act (Source: Martin Haimerl)

1.4.6 Conformity assessment of AI systems and products

The regulation of AI under the proposed AI Act is divided into the two phases of pre-market and post-market. Prior to the former, access to the European internal market for goods and services is granted on the basis of compliance with the essential requirements listed in the Act through a conformity assessment procedure. This applies in particular to those AI systems that fall into the "high-risk" category. Additional sectoral harmonization regulations of the EU (e.g. Machinery Directive, Medical Devices Regulation) may have to be observed and taken into account in the conformity assessment procedure. The classification into the category "high risk" and the associated requirement to carry out a conformity assessment procedure (Art. 16e) applies, according to Art. 6, both to safety components of products and to AI systems that constitute stand-alone products.

As part of the conformity assessment process, manufacturers of high-risk AI applications (as defined by the AI Act) should ensure or demonstrate that their products meet, in particular, the requirements for AI systems listed in Title III, Chapter 2. These include requirements in the areas of risk management (Art. 9), data and data governance (Art. 10), technical documentation (Art. 11), record-keeping requirements (Art. 12), transparency (Art. 13), human oversight (Art. 14), and accuracy, robustness, and cybersecurity (Art. 15). In addition, providers of high-risk systems must implement other obligations such as the establishment of a quality management system (Art. 17), registration obligations (Art. 51), post-market monitoring measures (Art. 61), and reporting of serious incidents and malfunctions (Art. 62).

Furthermore, high-risk AI systems as in Annex 3 must be registered in a publicly accessible European database for market access (Art. 16 (f), Art. 51). Post-market monitoring must be designed to actively and systematically collect, document, and analyze data on the performance of high-risk AI systems over their lifetime. This should be interpreted to mean that post-market monitoring is proportionate to the nature of the AI technology and the risks of the high-risk AI system (Art. 61). Overall, the risk management system and also human oversight must be implemented to accompany the AI system throughout its life cycle (Arts. 9 and 14). Furthermore, there must be a digital tracking of the functionality (Art. 12).

From a technical regulation perspective, the functionality of AI systems is supervised as part of post-market surveillance by national supervisory authorities in the individual European member states (Art. 59, 63) For individual sectors in the high-risk area, the authorities responsible under legal acts take over post-market surveillance. For financial transactions, for example, the Financial Supervisory Authority would be responsible. As part of post-market surveillance, access to data from data-driven models and product/system descriptive documents would be ensured by competent authorities (Art. 64), and the obligation to inspect AI systems would be imposed (Arts. 63-67).

According to Art. 43, there are two different ways of carrying out the conformity assessment procedure, which are shown in Figure 10. If there are corresponding harmonized standards that cover the fulfillment of the essential requirements of the planned AI Act and also the sectoral harmonization regulations, the manufacturer can refer to these standards. If this is the case and the manufacturer has fully applied the harmonized standards, they may carry out the procedure on the basis of an internal control using the abbreviated procedure described in Annex VI. After meeting the essential requirements, the high-risk AI system receives a CE marking (Art. 16 (i), Art. 19, Art. 49). These options apply to high-risk systems as defined in Art. 6, para. 2 and Annex III, i.e. systems for which classification as a high-risk product is not required by other sectoral harmonization legislation (as defined in Art. 6, para. 2 and Annex II). In the latter case, the implementation of the conformity procedure is carried out in conjunction with the respective sectoral regulations.

If either no harmonized standards are available or they are not or only partially applied, the conformity assessment procedure shall be carried out on the basis of the assessment of the quality management system and the assessment of the technical documentation with the involvement of a notified body. The associated requirements are listed in Annex VII. In this case, the quality management system is inspected by the notified body (Annex VII, item 3.2) and, in the case of a positive decision, is also inspected in the further course (Annex VII, items 3.3/3.4). In addition, the technical documentation of the respective AI system is inspected by the notified body. This must include unrestricted access to the test and training data via a suitable API (Annex VII, item 4.3) and, if necessary, access to the source code (Annex VII, item 4.5). In addition, it may require further testing to perform a proper conformity assessment (Annex VII, item 4.4). If the inspection is positive, the notified body issues an EU certificate on the assessment of the technical documentation (Annex VII, item 4.6). In addition, any change to the AI system that could affect its conformity with the requirements or its intended purpose requires the approval of the notified body that issued the aforementioned certificate for the AI system (Annex VII, item 4.7).

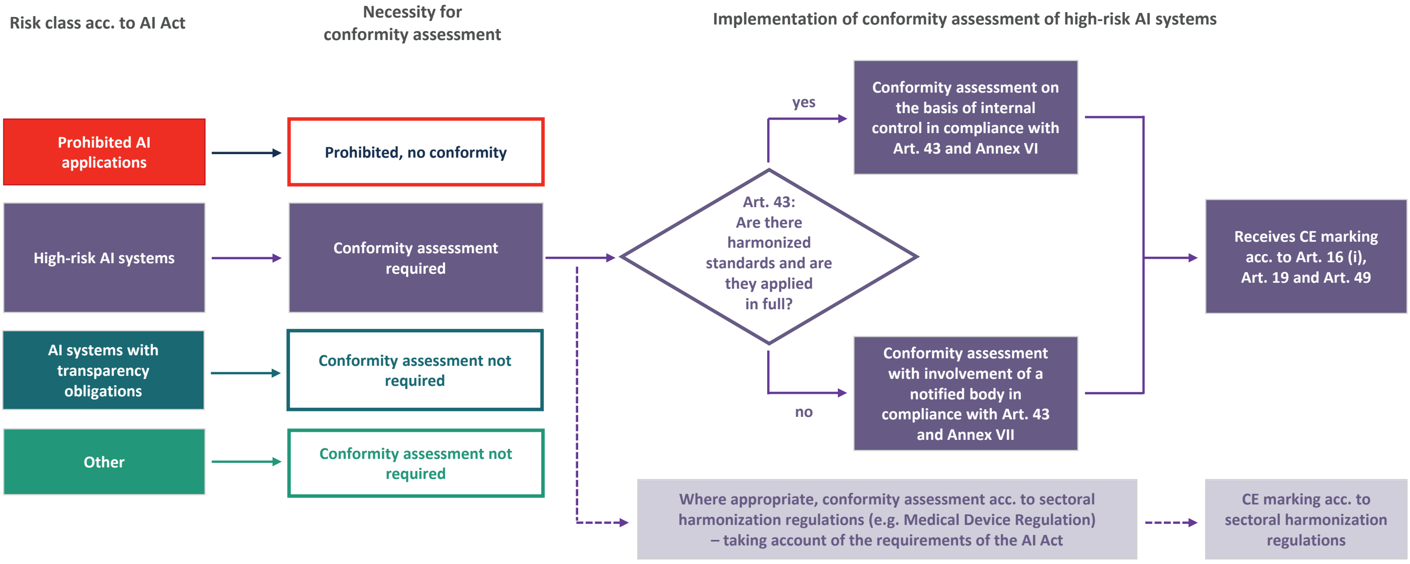

Risk class acc. to AI Act
Necessity for conformity assessment
Implementation of conformity assessment of high-risk AI systems
Prohibited AI applications
Prohibited, no conformity
High-risk AI systems
Conformity assessment required
AI systems with transparency obligations
Conformity assessment not required
Other
Conformity assessment not required
Art. 43: Are there harmonized standards and are they applied in full?
yes
no
Conformity assessment on the basis of internal control in compliance with Art. 43 and Annex VI
Conformity assessment with involvement of a notified body in compliance with Art. 43 and Annex VII
Receives CE marking acc. to Art. 16 (i), Art. 19 and Art. 49
Where appropriate, conformity assessment acc. to sectoral harmonization regulations (e.g. Medical Device Regulation) – taking account of the requirements of the AI Act
CE marking acc. to sectoral harmonization regulations

Figure 10: Variations of conformity assessment according to the AI Act (Source: Martin Haimerl)

1.4.7 Summary and discussion

In summary, from the point of view of those involved in the Standardization Roadmap, the proposal for European AI regulation currently under discussion has some key strengths despite its length and textual complexity: The planned AI Act is intended to create a uniform, common position in Europe in the field of AI, which will create transparency and legal certainty. This can improve and ideally also accelerate market introduction and control in the sense of value-based regulation. The risk-based approach aims to minimize threats while promoting innovation and market diffusion, especially for low-risk AI applications. The procedure for authorization or placing on the market can thus be adapted to the risk potential, although in the future there will be a need to look even more closely at the risk of specific products. The focus of the planned AI Act on the ethical values and legal foundations of the European Union, which are also to be implemented by AI applications based on the Act, is a special unique selling point of "AI Made in Europe" and can ideally ensure trust in the new AI technologies as well as promote their dissemination.

At the time of publication of this Standardization Roadmap, there are still some open questions, among other things, concerning the broad definition of AI in the Commission's draft. There is a need for a uniform description of the understanding of the term within the framework of the Act, so that currently existing inaccuracies in dealing with the AI term do not affect the implementation of AI-based products (e.g. in the required technical documentation and the planned European AI database). Overall, the requirements for managing high-risk AI systems still need some clarification and review. For example, products that have to comply with the AI Act as AI-based systems, as well as with sector-specific harmonization regulations, such as the Medical Device Regulation (MDR) in the case of medical devices, will result in increased efforts and expenses. This would be the case in particular if inconsistencies remained among the harmonization regulations or if different notified bodies had to be used, e.g. because the notified body previously commissioned by the company cannot cover both harmonization regulations (see Annex 13.1, Clause "Exemplary presentation using the example of medical devices" for a detailed presentation of this problem). This would have a particularly strong effect if insufficient capacities were built up at the notified bodies due to excessively short transition periods at the start of the AI Act's validity.

The same applies to the availability of harmonized standards. Through such standards, the approval processes can be implemented faster, more reliably and more uniformly. To this end, it is essential that the remaining ambiguities in the planned AI Act be eliminated and the processes relevant to standardization be initiated in good time. The experts involved must be given sufficient time to translate the complex requirements into practicable standards. The possibility that, in the absence of available harmonized standards, common specifications may take their place is seen as disadvantageous in comparison with the principle of standardization. Drawing on the expertise of representatives from the industry is a key factor in the successful implementation of regulatory requirements. This Standardization Roadmap shows what needs to be implemented to lay the groundwork.

1.5 Definition of AI

In the emergent general field of artificial intelligence, it is difficult to ensure a precise definition of terms due to a multitude of different perspectives and stakeholder backgrounds. The following key questions keep coming up in discussions about the core of "artificial intelligence":

a) Should the term refer to a scientific or a technical background?
b) Should the term refer to a system property or a system capability?
c) Should the term be limited to the function of AI systems or should it reference their implementation as well?
d) Should terms commonly associated with human intelligence (such as „knowledge" or „skills") be used to explain AI?

To make the terms clearer, a distinction is therefore often made between "AI systems" and "AI". Almost every organization that deals with artificial intelligence defines it differently, to a greater or lesser extent.

From the set of different definitions on the AI topic, two central definitions of "AI system" and "AI" will be highlighted here.

On the political level of social regulation, the European Commission's draft AI Act [4] should be mentioned here.

“AI system” here refers to software that generates results (for example, content, predictions, or decisions) according to human objectives using specific techniques or approaches, which in turn influence the context of the AI systems themselves.

The specific techniques and approaches are as follows:
a) Machine learning approaches including supervised machine learning, unsupervised machine learning, reinforcement learning, and a variety of methods including deep learning
b) Logic and knowledge-based approaches including knowledge representation, inductive (logic) programming, knowledge bases, inference engines and deductive engines, (symbolic) reasoning, and expert systems
c) Statistical models, Bayesian estimation, and search and optimization methods

“AI” in the proposed AI Act generally refers to the field of rapidly evolving AI systems technologies.

Meanwhile, at the international level of technical regulation, there is an International Standard on concepts and terminologies in AI (ISO/IEC 22989:2022) [16].

“AI systems” as in this standard means a constructed system that generates results (for example, content, predictions, recommendations, and decisions) according to human objectives. Four core features of AI systems are identified:
a) Interaction: the registration of information via sensors or human input
b) Context-sensitivity: some AI systems react to multiple sources of information
c) (Human) oversight: AI systems can operate with various degrees of human oversight and control
d) Adaptability: some AI systems are engineered to utilize dynamic data in real time and retrain to update their operation based on new data

“AI” is referred to here as the discipline of researching and developing mechanisms and applications of AI systems.

This Standardization Roadmap AI refers to the International Standard ISO/IEC 22989:2022 [16] for its definition of AI and AI systems.

2 Recommendations for action of the Standardization Roadmap AI

The aim of the Standardization Roadmap AI is to describe a framework for action that strengthens German industry and science in international competition for the best solutions and products in the field of artificial intelligence, and creates innovation-friendly framework conditions It is thus making a significant contribution to establishing "AI – Made in Germany" as a strong brand and developing new business models, disruptive innovations and scalable applications. German SMEs and the growing start-up scene in Germany in particular can benefit from this. Standards and specifications form the basis for technical sovereignty and create a framework that promotes transparency and provides orientation. Thus, they ensure security, quality and reliability and contribute significantly to the explainability of AI solutions – an essential basis when it comes to the acceptance of AI applications. The Standardization Roadmap AI offers great potential for both securing Germany's competitiveness and raising European value standards to the international level Not least for this reason, particular attention should be paid to the implementation of the Standardization Roadmap AI and its recommendations for action.

Recommendation 1: Develop, validate, and standardize a horizontal conformity assessment and certification programme for trustworthy AI systems
The EU Commission's current proposal for a European legal framework (AI Act) requires an application-agnostic, marketable conformity assessment and certification programme that makes the requirements of industry, public authorities, and civil society for AI systems objectively verifiable.

The lack of such a conformity assessment and certification programme threatens the economic growth and competitiveness of AI as a technology of the future. For example, statements about the trustworthiness of AI systems are not robust without high-quality testing methods, leaving the acceptance of AI systems in business and society unclear. The successful use of AI systems that meet the requirements of the European Act, and thus the European values, requires transparency throughout the supply chain in distributed and hybrid AI systems through informed, reliable, and reproducible testing of AI technologies.

The Standardization Roadmap AI therefore recommends the development, validation and standardization of an AI conformity and certification program as a top priority. Considerable preliminary work has already been done in the context of updating the Standardization Roadmap AI, so that implementation of this recommendation for action can begin immediately.

The fundamentals and architecture of the certification programme should be defined and harmonized by the standardization bodies. The harmonization concerns the scopes, the needs-based test criteria, the requirements and evidence, and the test methods for the certification of AI products, AI systems, and AI management systems. This would enable Germany to make a leading contribution to the development and standardization of an internationally recognized AI certification procedure.

To implement the recommendation, it is essential that the federal government fund and provide a budget for the following projects:

→ Develop and update an internationally accreditable AI certification procedure that fits into the existing certification infrastructure for products, services, processes and organizations.[10] In horizontal, application-agnostic standardization, the focus is on German projects already initiated for AI certification of products, hybrid systems, services and entire supply chains, and additionally management systems for organizations.

→ Initiate and implement research projects, especially in the field of high quality testing, and the reduction of unjustifiably great testing efforts. These include the following research areas:
  - uncertainty in neural networks
  - explainability and transparency
  - the development and certification of test tools for all test dimensions
  - the composition of test results

→ Transfer results into standardization and development of standards and specifications: In order to coordinate the results of the above-mentioned projects at national level and to introduce them promptly at the European and international levels, it is absolutely essential to finance corresponding standardization projects. In particular, experts must be recruited for this purpose and resources made available for their participation in the standardization bodies.

→ In order to develop AI certification and a corresponding infrastructure that fits into the testing and quality assurance of information technology systems as a whole, it is proposed that the management of the above-mentioned

[10] DIN EN ISO/IEC 17065:2013 [17] (in connection with DIN EN ISO/IEC 17067:2013[18] and the corresponding specifications ISO/IEC TR 17026:2015 [19] (products only); ISO/IEC TR 17028:2017 [20] (services); ISO/IEC TR 17032:2019 [21] (processes)), DIN EN ISO/IEC 17021-1:2015 [22] (organizations/management systems)

programme be jointly assigned to the national standards organizations and the Federal Office for Information Security.

Recommendation 2: Establish data infrastructures and elaborate data quality standards for the development and validation of AI systems

Data plays a central role in the realization of many AI systems, and the quality of AI systems often depends critically on data quality. Here, large amounts of data are needed both for training these systems and for validation (systematic testing). A prominent example is the development of large language models such as Open-GPT-3 or DALL-E 2, which require datasets with several 100 million training data. In addition to training appropriate systems, data is also needed for the systematic testing of AI systems. Many test scenarios are needed especially for the validation of AI systems operating in an open-world context. The availability of corresponding datasets is thus also a strategic success and competitive factor for the German AI industry and start-ups in particular. This requires appropriate data infrastructures that collect, curate, describe through appropriate metadata, and make available suitable datasets. In the generation of such datasets, synthetic data in particular also plays a decisive role, since for some AI applications there is not enough real data available or some test scenarios occur too rarely for the availability of real data to be sufficient for adequate validation. Depending on the intended use, both the provision of open source datasets and marketplaces that enable trading with corresponding data are conceivable. Regulatory data infrastructures may also be required for the approval of AI systems with a critical context of use – for example, in the medical sector – which provide datasets for the approval of these AI applications.

The realization of such data infrastructures should rely on current data architectures such as data meshes, use data virtualization techniques and, where possible, build on existing structures such as Gaia-X or the European Health Data Space. At the same time, appropriate tools need to be developed that quality check datasets and identify subsets of the data on which the corresponding AI systems perform less well and can be used to generate targeted high-quality synthetic data.

Standards and specifications are of particular importance in the provision of such datasets and their data infrastructures to ensure interoperability while defining quality standards. Appropriate data quality standards can ensure that datasets are representative, complete, error-free, and balanced, for example.

The Standardization Roadmap AI therefore recommends the promotion of such data infrastructures by the public sector and, at the same time, the support of standards organizations in the development of corresponding data quality standards. Since the successful development of standards and specifications depends to a large extent on the participation of relevant experts, the provision of the necessary resources for participation in the standardization bodies must be ensured by the German government.

The activities that focus on the validation of AI systems should be closely connected with the certification and conformity assessment programme (see Recommendation for action 1).

Recommendation 3: Consider humans as part of the system in all phases of the AI life cycle

The current draft of the European Artificial Intelligence Regulation (AI Act) makes extensive demands on high-risk systems in particular for the involvement of humans, e.g. transparency for those affected and involved, human oversight in different roles and intervention options, right up to a “stop button” triggered by humans. Which transparency is sufficient in which context for which target group, how human oversight should be implemented, and which basic information must be available as a basis for human intervention in the system – these are all questions that must be thought of from the human perspective and according to which the technical and social components must be developed and aligned.

To implement these sociotechnical aspects in AI systems, the following challenges need to be addressed:

- → Suitability: Technical components are to be selected based on sociotechnical requirements.
- → Participation: The definition and selection of relevant actors who should be involved must be operationalized.
- → Ethics: Social and ethical issues must be operationalized with the help of established models, anchored in a measurable way as early as the development of the technology, and based on the latest research on discrimination sensitivity.
- → Culture: An adequate organizational culture must be established (in the work context, e.g., the corporate culture), because this must also be co-developed during AI deployment. To this end, the relevant actors must be sensitized, qualified and involved in the process through appropriate change management.
- → Tools: Across the life cycle of an AI system, humans need to be supported with processes, methods, and tools –

from goal setting to development to operation with iterations and re-validation.

Research projects have already produced findings on some of the above-mentioned challenges. These must now be adapted and fine tuned to meet the possibly still changing requirements of the planned AI Act.

This results in the following specific recommendations for action:

→ To funding agencies: "Lighthouse" projects should concretely test how the involvement of affected people and the people involved in all phases of the AI life cycle can succeed in different contexts.

→ To standardization bodies: Accompanying this, the necessary standards for the sociotechnical aspects of the proposed AI Act must be developed in a timely manner, particularly on human oversight and the necessary transparency requirements.

→ To standardization organizations and policymakers: In order to do justice to the far-reaching social responsibility, it is necessary to pay particular attention to the balanced participation of all relevant target groups in the standardization organizations and to actively promote this (e.g. science or civil society).

→ To policymakers: For the sociotechnical perspective, which has so far been underrepresented in standardization, it is also imperative to recruit experts and make their capacities available in standardization bodies in order to productively bundle findings at the national level and contribute them at the European and international levels.

Recommendation 4: Develop specifications for the conformity assessment of continual or incremental learning systems in the field of medicine

AI systems can be continuously improved with more data and information. This results in significant potential for the improvement of AI systems already in use in the field, since, for example, new training data as well as information on faulty behaviour and corrections can be obtained. On the other hand, the integration of the new data must be implemented at a high level of quality and underpinned by appropriate testing processes in order to meet the high safety requirements in the medical sector.

In this context, there are several challenges with two fundamentally different approaches:

On the one hand, in the sense of a continuously learning system, data can be collected locally and entered into the model/system in order to improve it and/or adapt it specifically to local regions, individual hospitals, or individual patients (groups). On the other hand, updates can be made in stages as part of a simplified recertification process or conformity assessment procedure.

Specifically, policy and standardization requirements need to be developed for the following aspects.

→ It cannot be determined in advance which and accordingly how much data is required for continual/incremental learning. Theoretically, all data that has flowed into the online learning databasecan be of significance. This data cannot usually be made readily available. Therefore, it must be clarified how the management and dissemination of the data as well as the application of the AI system is to be designed in order to make the changing AI system traceable and thus also auditable.

→ A validation process is needed that is to be defined based on concrete, medicine-specific protection targets and that can comprehensively and reliably verify a model update. This includes the definition and testing of requirements for both the validation process and the AI system (especially the updated system). This may result in the validation process itself being subject to validation.

→ An „agile release/conformity assessment process" is needed that implements the clinical validation required under the Medical Device Regulation (MDR) of AI systems improved in parts by online-acquired data in such a way that recertification or a new conformity assessment procedure of the entire system does not have to be performed each time.

For the release or conformity (re-)assessment of such systems, recognized test requirements and test procedures are lacking, especially in Europe. Ultimately, these test requirements and test methods can only be successfully brought to market if the actors from politics, standardization, research and industry support them. In order to achieve certification or market access, boundary conditions must be specified in advance that allow the automated release of continual or incremental learning systems.

This results in the following specific recommendations for action:

→ A "lighthouse" project is to be initiated and carried out that is funded by the public sector (e.g. the ministries

BMBF, BMWK, BMG and others) and that looks at various domains or aspects:

- analyses in medical imaging,
- oncology/cancer detection,
- automated intensive care,
- identification and therapy of sepsis.

→ Each of these domain-specific subprojects requires collaboration between university and non-university research institutions, real-world hospitals, medical device manufacturers, tech/IT companies, TIC companies (TIC: testing, inspection, certification), and standards organizations, and must be provided with an appropriate budget and a cross-project governance structure that ensures coordination of content.

→ In addition to a scientific advisory board, the establishment of a project-office office is recommended which will be responsible for the translation of project results into standards, specifications and generally practiced test methods and their cross-industry utilization and international placement. Such a project-office should be launched as soon as possible.

Recommendation 5: Develop and deploy secure and trusted AI applications in mobility through best practices and assurance

The use of AI technologies in the context of mobility is characterized by complex boundary conditions. These are characterized by complex decision and control systems, which interact in a sensorimotor loop in a constantly changing environment both with the environment itself and with a wealth of other actors – in combination with the high risks of malfunctions for humans and the environment. Standards and specifications for the dynamic type approval of mobility systems whose functionality is based at least in part on the use of AI technology are therefore urgently needed in order to enable or guarantee sufficient performance on a sustained basis under these complex boundary conditions, as well as the necessary trustworthiness and safety. While the various aspects of trustworthiness are already largely specified by the draft AI Act, operationalizing these aspects requires concretization across the entire life cycle of an AI system. In particular, these standards and specifications should ensure that ...

→ the efficient (further) development, validation, successive introduction and continuous assurance in operation are supported by a best practice catalogue that ensures the performance and trustworthiness of the systems. These measures should include qualified procedures and tools for development and testing, as well as explainable AI procedures whose relevant properties can be analyzed and tested to demonstrate security and trustworthiness.

→ the evidence and validations of trustworthiness and security to an independent third party are enabled. In this context, standards and specifications for minimum requirements (in particular, intolerable residual risks to safety) as well as the other essential trust aspects (including IT security, robustness, transparency, traceability, data protection and non-discrimination) must be defined and taken into account over the entire life cycle of an AI system, with the required safeguarding being designed on the one hand as a function of the risk class of the specific mobility application, and on the other hand of the respective „socially accepted residual risk“.

To enable the successive deployment of secure and trustworthy AI-based mobility applications despite remaining and emerging uncertainties, agile approaches to regulation (see [23]) and standardization are needed to enable continuous monitoring and adjustment of the effectiveness of regulatory levers. This requires monitoring of operational risk factors and presupposes certain societal expectations, as well as a close integration of standardization and regulation.

Recommendation 6: Develop overarching data standards and dynamic modelling techniques for the efficient and sustainable design of AI systems

AI systems are increasingly being used to address issues of relevance to the present day. This concerns the intelligent control of systems and the formulation of recommendations for action across sectors. There is a high degree of interdisciplinarity here, in that data domains that were previously clearly separable are being merged and static modelling standards are being made more flexible. The goal-oriented design of new data standards and modelling procedures relies both on standards for interpreting and aggregating data and on procedural standards. Normative definitions represent cross-industry guidelines for industry and science that can provide answers to the following questions:

→ How should data syntax and semantics and the AI systems based on them be designed so that (partially) autonomous systems can be operated efficiently and resiliently?

→ How can common regulatory frameworks or frameworks for data sources from different sectors be designed to facilitate continuous data acquisition and communication for and in AI as well as ML?

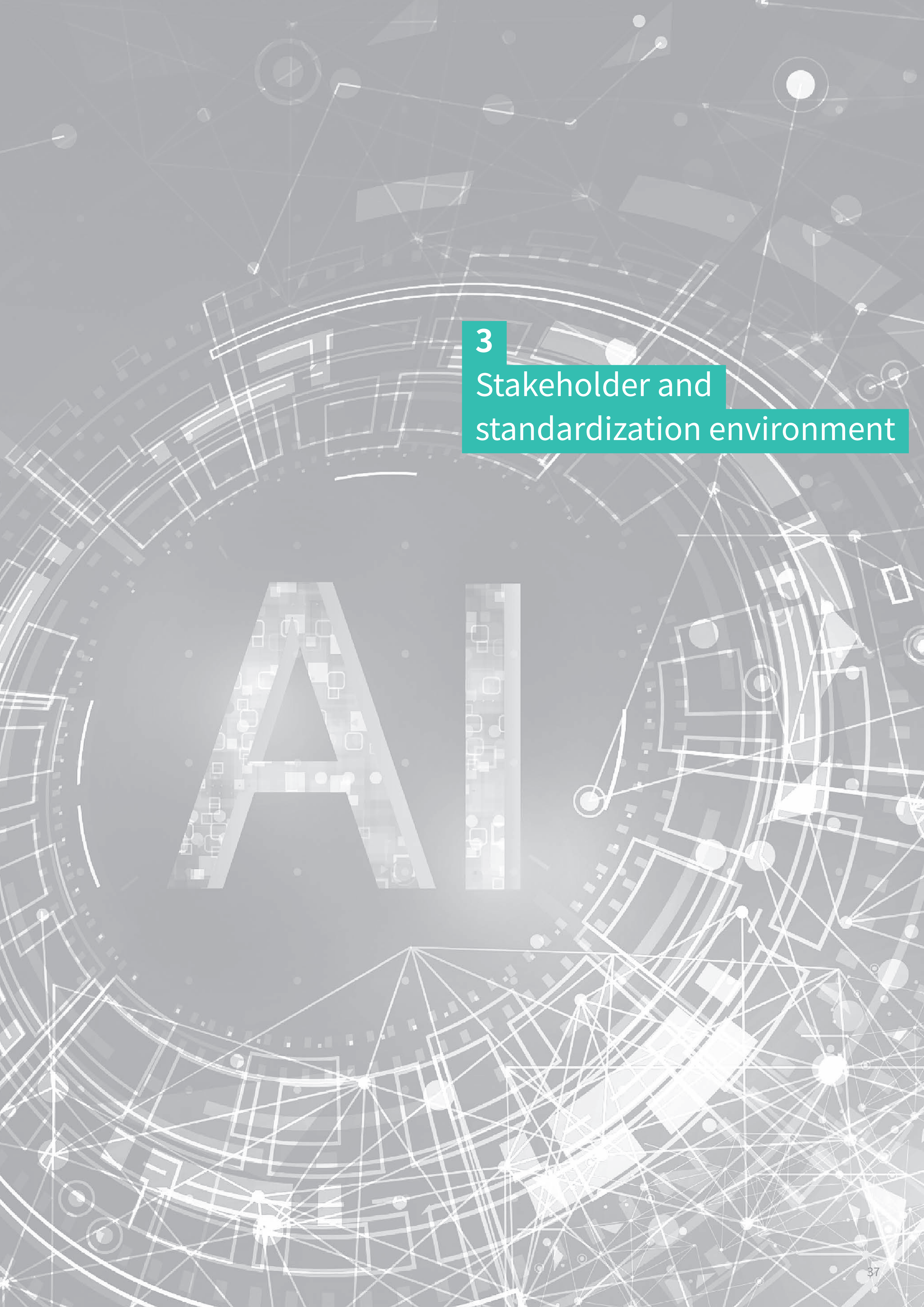
→ How is the efficiency and sustainability quality of AI to be operationalized and evaluated?

→ To what extent can the ongoing, agile development of new domain-specific data standards be accommodated in the standardization of overarching frameworks?
→ In this context, how can we ensure that data with higher temporal and geographic resolution are given sufficient consideration?

Consequently, there is a comprehensive need for standardization to overcome data system boundaries and to develop reference procedures. These standardization needs can be met by the joint action of actors from standardization, industry and science. Therefore, pilot projects in the public funding context are recommended on the following aspects:

→ Establishing a common terminology, semantics, taxonomy, and the data mappings and schemas based on them in the domains of materials science and construction to determine energy efficiency and environmental impacts, among other things, for building ESG datasets and using them in AI-based planning tools for future resource consumption. Such a standardization project can involve stakeholders from municipal construction planning, materials management, finance, and research on energy efficiency in the materials and building sector, can be included in funding programmes of the Ministry for Economic Affairs and Climate Action (BMWK) or Ministry for the Environment, Nature Conservation, Nuclear Safety and Consumer Protection (BMUV), and should be initiated in the first half of 2023.
→ Developing an industry-independent communication format for determining the energy and resource consumption of goods and services. Such a standardization or pilot project is of interest to actors from all sectors or industries with reference to private customers and to socio-ecological research, can be placed in the funding context of the Ministry for the Environment, Nature Conservation, Nuclear Safety and Consumer Protection (BMUV) and should be initiated at the beginning of 2023 in view of the range of stakeholders and the corresponding need for coordination.
→ Developing a methodology to assess the runtime, accuracy, and sustainability performance of AI and ML systems. A corresponding standardization project should involve representatives of all industries with current and near-future predicted intensive AI and ML use, AI certifiers, and researchers with expertise in algorithms and machine learning, should be placed in the public funding context, and, given the range of possible participants, should be initiated soon (early 2023) in the participant acquisition process.
→ Establishing a synergistic dynamic modelling approach for reference architecture models in smart manufacturing and the smart grid for mapping dynamic variable behaviour and identifying critical system areas. This should be based on the RAMI 4.0 (Reference Architecture Model Industrie 4.0) and the SGAM (Smart Grid Architecture Model) Reference Designation Models, which support and simplify the development of sample solutions as a „systems of systems“ approach.
→ Developing a dynamic calculation method for CO_2 emissions from the electricity mix to account for the geographic-temporal volatility of sustainable electricity generation. Such a pilot or standardization project should involve stakeholders from the electricity industry and geodesy or cartography, be located in the funding context of the Ministry for Economic Affairs and Climate Action (BMWK) or Ministry for the Environment, Nature Conservation, Nuclear Safety and Consumer Protection (BMUV) and should be initiated in the first half of 2023.

3 Stakeholder and standardization environment



Numerous stakeholders, initiatives, committees, and standardization activities at the national, European, and international levels are intensively dealing with the topic of AI. The following chapter presents a selection of the most important stakeholders and initiatives in the AI environment.[11]

3.1 Innovative political initiatives

Plattform Lernende Systeme (PLS) (Platform Learning Systems)

The Platform Learning Systems (PLS)[12] was initiated in 2017 by the German Federal Ministry of Education and Research with the aim of shaping AI for the benefit of society and developing recommendations for action for the responsible use of AI. In seven thematic working groups, the initiative pools the knowledge of around 200 experts from science and industry, as well as decision-makers in the innovation ecosystem and in politics, on topics such as the legal and social framework conditions for the application of AI.

The working groups are:

- → Technologies and data science
- → Work and skilling, human-machine Interaction
- → IT security, privacy, law and ethics
- → Business model innovations
- → Mobility and intelligent traffic systems
- → Health, medicine and care
- → Learning robotic systems

The Platform Learning Systems also provides an overview of Germany as a centre for AI: For example, the AI map[13] shows AI applications, research institutions, transfer centres and study programmes throughout Germany. The "AI monitoring"[14] uses various indicators to illustrate the status quo as well as development potential in research and transfer.

AI Competence Centres

A central component of the German AI cosmos and the German government's AI Strategy are the National Competence Centres for AI Research. Since July 2022, six Competence Centres have received permanent funding from the Federal Ministry of Education and Research (BMBF) and are to conduct nationally and internationally networked cutting-edge research and establish and expand AI competencies in Germany. The overriding goal is to secure Germany's technological sovereignty in artificial intelligence. The Competence Centres are designed to enable scientific breakthroughs, spawn new start-ups and business models, accelerate research transfer, train AI professionals, and create new jobs.

The six AI Competence Centres include:

- → **Berlin Institute for the Foundation of Learning and Data (BIFOLD)**
  Funded by the BMBF and the Berlin Senate Chancellery for Science and Research, BIFOLD[15] is an association of research institutions that focus on big data management and machine learning (ML). Specifically, the research initiative contributes to the development of tools and infrastructures for AI applications. A large number of the tools developed are offered by the initiative as open source[16] tools.
- → **German Research Centre for Artificial Intelligence (DFKI)**
  As the world's largest independent research centre for AI, the German Research Centre for Artificial Intelligence (DFKI)[17] conducts research at several locations in Germany on solutions for human-centred use of artificial intelligence. It focuses in particular on challenges facing society as a whole, such as human-made climate change, social injustice, and the fight against disease, and initiates, implements, and supports numerous activities to place reliable and trustworthy AI from Germany and Europe at the forefront of international competition.
- → **Munich Centre for Machine Learning (MCML)**
  A joint research initiative of the Ludwig-Maximilians-Universität München (LMU) and Technische Universität München (TUM), the MCML[18] receives funding as part of the German and Bavarian AI Strategies. The association consists of 50 research groups, which focus on basic research, as well as research on applied ML. To ensure the transfer of findings to industry, the initiative supports the training of students and provides training in AI applications for industrial companies.

[11] The presentation makes no claim to completeness.

[12] https://www.plattform-lernende-systeme.de/startseite.html

[13] https://www.plattform-lernende-systeme.de/ki-in-deutschland.html

[14] http://www.kimonitoring.de/

[15] https://bifold.berlin/

[16] https://www.bifold.berlin/impact-transfer/open-source-systems-tools-data

[17] https://www.dfki.de/web

[18] https://mcml.ai/

→ **Lamarr Institute for Machine Learning and Artificial Intelligence**
Within the framework of the German AI Strategy, the Lamarr Institute[19] is funded by the BMBF and the state of North Rhine-Westphalia. The institute is the result of an initiative of the Technical University of Dortmund, Fraunhofer IML, Fraunhofer IAIS and the University of Bonn and replaces the Machine Learning Competence Centre Rhine-Ruhr (ML2R)[20]. In addition to research on ML technologies, the initiative also offers educational opportunities for pupils, students and young scientists. The initiative pays particular attention to sustainable innovations and social justice.

→ **Centre for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI)**
The Centre for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI)[21] focuses their research on applied AI and AI methods in the area of big data as well as data analytics. Research is conducted in the three core areas of "Applied AI & Big Data", "AI Algorithms & Methods" and "Big Data Analytics & Engineering", whereby the ethical and social dimensions as well as security and scalability are considered in all areas. As an association of 13 research institutions, the initiative is funded by the BMBF and the federal state of Saxony and has set itself the goal of ensuring the transfer of scientific results through cooperative projects with industrial partners.

→ **Tübingen AI Centre**
The Max Planck Institute for Intelligent Systems and the Eberhard Karls University of Tübingen have joined forces in the Tübingen AI Centre[22] to develop learning systems that have a positive impact on society. The initiative is funded by the BMBF and the Ministry of Science, Research and the Arts of the state of Baden-Württemberg (MWK BW). Due to the close proximity to the Cyber Valley initiative, a partnership has been formed between the two initiatives.

Mittelstand-Digital
The Mittelstand-Digital[23] initiative is an initiative of the BMWK, where companies from all over Germany can obtain information and further training on topics related to digitalization. Specifically, the initiative offers digital learning opportunities as well as practical examples and demonstration sites to see digital technologies in action. The initiative also spans a nationwide network of centres to provide companies with a locally available source of information. The "Mittelstand-Digital" strategy also includes AI trainers who provide information and advice on the topic in workshops, lectures and roadshows.

Mittelstand-Digital Centres and Mittelstand-4.0 Competence Centres
The (competence) centres[24] funded by the BMWK serve as the first point of contact for companies seeking information on the topic of digitalization. A total of 66 competence centres are currently being funded for this purpose. The range of services includes clarification of questions and training in the safe use of new technologies, support in testing developed applications, as well as advice on IT law and the development of digital business models. To learn about centre locations, BMWK maintains an interactive map on its website.[25]

Competence Platform KI.NRW
With the competence platform KI.NRW[26], the state of North Rhine-Westphalia (NRW) has created a central contact point for artificial intelligence in NRW, which is intended to accelerate the transfer of AI from cutting-edge research to industry. In addition to knowledge transfer, the competence platform promotes AI projects for establishing AI technologies in the broad industry and for professional qualification. The aim is to ensure and support efficient technology transfer and close cooperation between SMEs, start-ups, universities, colleges and research institutions in NRW. In addition to this, the platform also focuses on social aspects and ethical principles for the design of artificial intelligence.

[19] https://lamarr-institute.org/

[20] https://www.ml2r.de/

[21] https://scads.ai/

[22] https://tuebingen.ai/

[23] https://www.mittelstand-digital.de/MD/Navigation/DE/Home/home.html)

[24] https://www.mittelstand-digital.de/MD/Redaktion/DE/Artikel/Mittelstand-4-0/mittelstand-40-kompetenzzentren.html

[25] https://www.mittelstand-digital.de/MD/Navigation/Karte/SiteGlobals/Forms/Formulare/karte-formular.html

[26] https://www.ki.nrw/

Regional Competence Centres for Labour Research

Within “Future of Work: Regional Competence Centres for Labour Research”, further competence centres are being funded by the BMBF with the aim of linking labour research more closely with work design in company practice and university education, and strengthening the transfer of new findings to society. These centres include:

→ **Künstlich und Menschlich Intelligent (K-M-I)**
This competence centre researches the use of artificial intelligence in the field of work design, for example by supporting intelligent assistance systems in production planning and control or in the maintenance and servicing of complex facilities. Researchers will investigate the potential of intelligent technical systems in collaboration between humans and machines from an occupational science perspective and test them in companies.

→ **WIRKsam**
The funded competence centre WIRKsam[27] is an initiative that researches AI innovations in work and process flows. In addition to developing new concepts, the initiative also looks at operational implementation. Companies from the coal and textile region of the Rhineland are the main target group. The competence centre is funded by the BMBF until October 2026 and is supervised by the Karlsruhe project management organization.

Competence Centre for AI Systems Engineering (CC-KING)

The Competence Centre for AI Systems Engineering (CC-KING)[28] was established by three research institutions from Karlsruhe (Fraunhofer IOSB, Forschungszentrum Informatik (FZI) and Karlsruhe Institute of Technology (KIT)). Funded by the Ministry of Economics, Labour and Tourism Baden-Württemberg (WM BW), the competence centre aims to facilitate the use of artificial intelligence (AI) and machine learning (ML) methods in engineering from a practical perspective. With a focus on industrial, sustainable production and demand-driven mobility, basic research is conducted here and methods are developed to improve operational work.

Observatory on Artificial Intelligence in Work and Society

Launched by the think tank of the Federal Ministry of Labour and Social Affairs (BMAS), the AI Observatory[29] focuses on the five fields of action: Technology foresight and technological impact assessment, AI in labour and social administration, Regulatory framework for AI and social technology design, Building international structures and European networking, and Social dialogue and networking. The aim of the Observatory is to provide an overview of developments in AI and to assess and positively influence their impact on society.

Cyber Valley

Founded by the state of Baden-Württemberg and research institutions and commercial enterprises, Cyber Valley[30] is an initiative to strengthen research in ML, robotics and computer vision. In addition to the expansion of basic research, Cyber Valley also offers the possibility of funding start-ups that bring the scientific knowledge gained into commercial application. Cyber Valley concentrates its support primarily in the Stuttgart-Tübingen area.

AI Quality & Testing Hub

The concept of an AI Quality & Testing Hub[31] has been driven forward by VDE and TÜV in cooperation with several federal states since 2020. It is intended as an institution with European reach that brings together all the pieces of the puzzle necessary for the assessment and management of AI quality, e.g. an overview of the state of research, access and development of training datasets, simulation environments with standardized interfaces, training and competence acquisition, as well as tailored quality improvement for manufacturers and users/operators of AI products.

AI4Germany

AI4Germany[32] is an umbrella initiative for the implementation-oriented promotion and implementation of AI. Founded by the Munich Start-up Accelerator TUM, the initiative sees itself as an application-oriented supplement to PLS and AI4Europe. The goal of the alliance is to strengthen Germany’s position as a location for developing high-tech AI applications.

[27] https://www.arbeitswissenschaft.net/wirksamweb/

[28] https://www.ai-engineering.eu/

[29] https://www.ki-observatorium.de/

[30] https://cyber-valley.de/de

[31] https://www.vde.com/de/presse/pressemitteilungen/ai-quality-testing-hub

[32] https://www.ai4germany.de/

Initiative for Applied Artificial Intelligence

59 partners from industry, the public sector and research have joined forces in the Initiative for Applied Artificial Intelligence (appliedAI)[33]. Aiming to create a collaborative platform that trains people and drives innovation, appliedAI is Europe's largest initiative. The platform's services range from consulting and training to exchanges and lectures to access AI tools/ecosystems and start-ups. AppliedAI is also a member of AI4Germany.

Platform Industrie 4.0

The Platform Industrie 4.0[34] is a network for promoting the digitalization of industry. It is led by the Federal Ministry for Economic Affairs and Climate Action (BMWK) and the Federal Ministry of Education and Research (BMBF) together with technology companies, associations and research organizations. In addition to topics such as geopolitical crises and supply chain resilience, the network regularly addresses AI, for example as a focus topic in the "Technology and Use Cases" working group.

KI-Transfer-Hub SH

The state of Schleswig-Holstein launched its own initiative KI-Transfer-Hub SH[35]. With this initiative, the state aims to enable SMEs and start-ups in particular to incorporate AI technologies into their business models. Partners from science and industry in northern Germany are providing support. The European Union is supporting this initiative with funding from the European Regional Development Fund.

In addition to the aforementioned initiatives, there are other associations beyond Germany's borders that aim to promote AI at the European and international level, including:

CLAIRE

The Confederation of Laboratories for Artificial Intelligence Research in Europe (CLAIRE)[36] is a pan-European alliance of 445 research institutions that aims to strengthen research and innovation in AI. As a partner of HumaneAI, CLAIRE contributes to the development of trustworthy AI. Since its establishment in The Hague in 2018, additional branches have been added in Germany, Norway, the Czech Republic, Italy, Switzerland and Belgium. The initiative's scope of work includes machine learning, knowledge representation and reasoning, natural language processing, as well as topics such as robotics, computer vision, and ethical and social aspects.

ELLIS

The European Laboratory for Learning and Intelligent Systems (ELLIS)[37] is a European network that was founded in 2018. ELLIS sees itself as a driver to strengthen Europe's economic position in AI development. Therefore, in addition to basic research, it is also driving the creation of new AI start-ups. There is a close relationship between ELLIS and CLAIRE, as both initiatives complement each other's efforts ELLIS is now represented at 35 locations in Europe, eight of which are in Germany.

AI4EUROPE

As a successor to AI4EU, AI4EUROPE[38] was launched in 2022 and provides a platform for research groups to share scientific knowledge and thereby drive further innovation. The platform was launched at the University College Cork in Ireland. In addition to research and education, the platform also provides an opportunity for industry to share their use cases.

I-DAIR

The International Digital Health & AI Research Collaborative (I-DAIR)[39] is an initiative that aims to improve research in digital health and AI in healthcare. In this way, the digital transformation will be used to help all countries and communities achieve an improved quality of life. This association of international research institutes now networks more than 40 partners. Two standout projects from the platform are the Global Research Map of Digital Health and AI[40], and the Real Time Epidemiology & Dashboard[41].

[33] https://www.appliedai.de/de/

[34] https://www.plattform-i40.de/IP/Navigation/DE/Home/home.html

[35] https://kuenstliche-intelligenz.sh/de/startseite

[36] https://claire-ai.org/?lang=de

[37] https://ellis.eu/

[38] https://www.ai4europe.eu/

[39] https://www.i-dair.org/

[40] https://grm.i-dair.org/

[41] https://www.i-dair.org/pathfinder/rted

3.2 Standardization environment

The contents of valid standards and specifications represent the current state of the art in science and technology. Each standardization document represents the essential properties (e.g. of a product), requirements (e.g. for a service) or procedures (e.g. of processes), which are usually developed on a consensual basis by participants from the stakeholders (industry, science, research, users, consumer protection, occupational health and safety, trade unions, public authorities and environmental protection).

By specifying technical and compatibility requirements for products, services or processes, but also by defining terms or interfaces, interoperability is ensured and the protection of people, the environment and things is ensured. In this way, standards and specifications create transparency and trust in new applications and technologies.

The need for a new standard or specification is often recognized by these stakeholders. In principle, however, anyone can apply to the responsible national standards institute (DIN for Germany) for the preparation of a standard.

Depending on the type of content, target group and global economic relevance, standardization work is carried out at national, European or international level (see Figure 11). Although there are differences between these three levels, they all have one thing in common; the standardization work is carried out by experts who are sent by their national standards institute to work at the European level (in the case of CEN/CENELEC) or international level (in the case of ISO/IEC). The national standards institutes of the participating countries thus represent the link between those stakeholders with know-how and the active development of standards.

As the responsible standardization organization of the Federal Republic of Germany, DIN[42] has represented German interests in European standardization (at CEN[43]) and in international standardization (at ISO[44]) since 1975. Standardization in the fields of electrical engineering, electronics and information technology is handled nationally and internationally by the

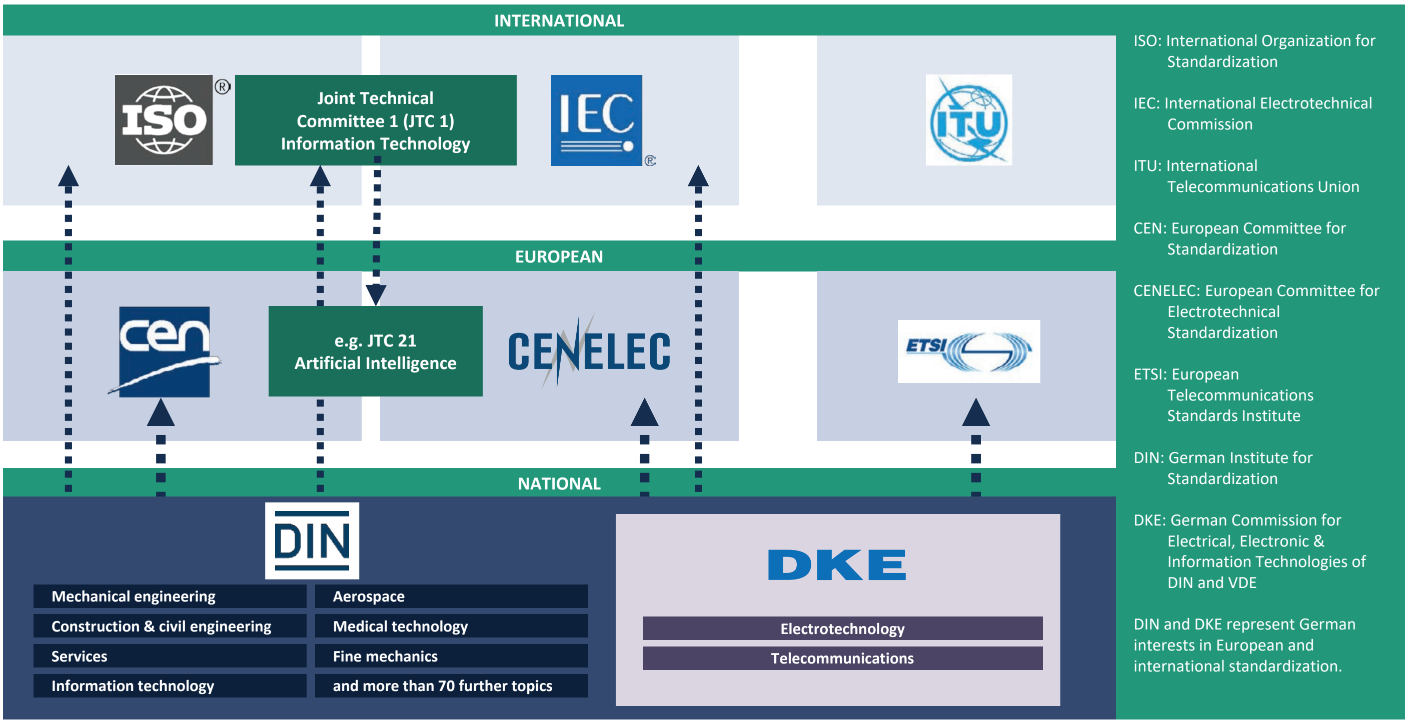

INTERNATIONAL
ISO
Joint Technical Committee 1 (JTC 1) Information Technology
IEC
ITU
EUROPEAN
cen
e.g. JTC 21 Artificial Intelligence
CENELEC
ETSI
NATIONAL
DIN
Mechanical engineering
Aerospace
Construction & civil engineering
Medical technology
Services
Fine mechanics
Information technology
and more than 70 further topics
DKE
Electrotechnology
Telecommunications
ISO: International Organization for Standardization
IEC: International Electrotechnical Commission
ITU: International Telecommunications Union
CEN: European Committee for Standardization
CENELEC: European Committee for Electrotechnical Standardization
ETSI: European Telecommunications Standards Institute
DIN: German Institute for Standardization
DKE: German Commission for Electrical, Electronic & Information Technologies of DIN and VDE
DIN and DKE represent German interests in European and international standardization.

Figure 11: Levels of standardization work (Source: DIN)

[42] German Institute for Standardization, www.din.de

[43] Comité Européen de Normalisation, European Committee for Standardization, https://www.cencenelec.eu/

[44] International Organization for Standardization, www.iso.org

DKE[45], which represents German interests at both CENELEC[46] and IEC[47]. Today, standardization work is concentrated at European and international level, with the responsibility within Germany lying with DIN and DKE, which coordinate the national work and contribute the German voice at the European and international levels through delegates and experts.

Standards are developed according to established principles at national, European and international level, taking into account procedural and design rules. In work in committees, the current state of the art is recorded by representatives of all stakeholders (e.g. manufacturers, consumers, trade, universities, research institutes, authorities, test institutes, etc.). Standards are created by consensus of all stakeholders.

“Specifications” are documents such as Technical Reports (TR), pre-standards, specifications (TS, DIN SPEC), consortium standards, application rules (AR), guidelines, expert recommendations, etc. These are often created for topics with a low level of maturity which may not yet be fully established in the market. The preparation and publication is carried out by the standardization institutes and other organizations and technical rule makers. Full consensus and the involvement of all stakeholders are not mandatory in the development of specifications.

3.2.1 AI standardization at national level

Standardization work on artificial intelligence is currently taking place at all three levels (national, European and international). At the national level, special mention should be made of the **DIN/DKE Joint Committee “Artificial Intelligence” NA 043-01-42 GA**[48], which was initially established by DIN in 2017 and further developed into the DIN/DKE Joint Committee at the end of 2021. More than 80 experts from business, science, politics and civil society are involved in the committee and develop standards for tools, processes and fields of application of artificial intelligence, always taking into account social opportunities and risks.

As the national mirror committee, the Joint Committee is responsible for consolidating German opinion and sends the German delegation to both the European standardization body (CEN/CENELEC/JTC21) and the international standardization body (ISO/IEC/JTC1).

It is one of the most important AI-relevant bodies for the implementation of European requirements (from Regulations, the AI Act, etc.) and plays an important role in the development of corresponding standards.

Figure 12 shows the structure of the German national AI joint committee.

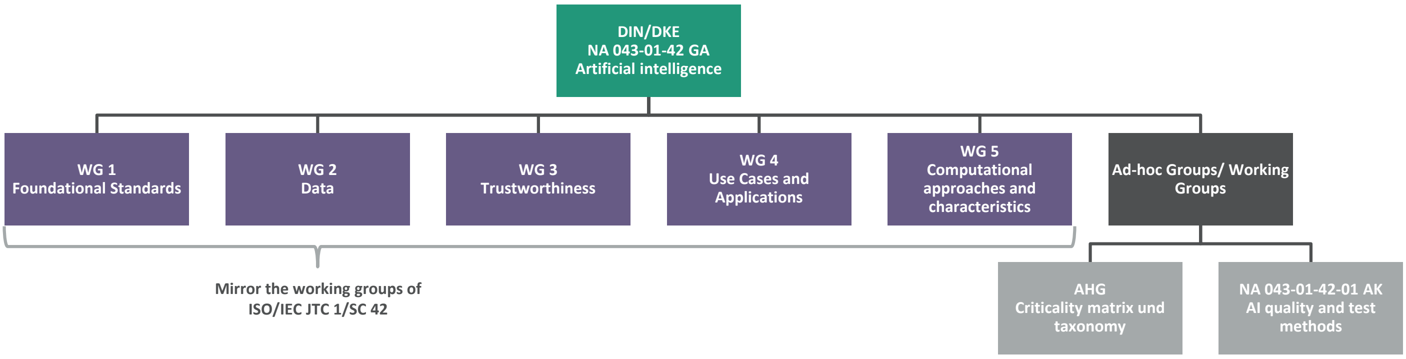

DIN/DKE
NA 043-01-42 GA
Artificial intelligence
WG 1
Foundational Standards
WG 2
Data
WG 3
Trustworthiness
WG 4
Use Cases and Applications
WG 5
Computational approaches and characteristics
Ad-hoc Groups/ Working Groups
AHG
Criticality matrix und taxonomy
NA 043-01-42-01 AK
AI quality and test methods
Mirror the working groups of
ISO/IEC JTC 1/SC 42

Figure 12: Structure of the national Joint Committee on AI (Source: DIN)

[45] DKE German Commission for Electrical, Electronic & Information Technologies of DIN and VDE www.dke.de

[46] Comité Européen de Normalisation Électrotechnique, European Committee for Electrotechnical Standardization, www.cenelec.eu

[47] International Electrotechnical Commission, www.iec.ch

[48] See https://www.din.de/de/interdisziplinaerer-arbeitsausschuss-zu-kuenstlicher-intelligenz-826618

In addition, the Working Committee "Information Security, Cybersecurity and Privacy Protection" NA 043-04-27 AA[49] should also be mentioned, whose topics are of particular relevance to AI.

3.2.2 AI standardization at European level

Since 2019, the topic of AI has been of great interest to European policymakers. With the New Legislative Framework, Europe has long had a unique and proven mechanism for the interaction of standardization and regulation, which is now to be applied to the subject of AI on the basis of the draft AI Act (see Chapter 1.4). At European level, the central task of standardization is to deal with Europe-specific aspects and to support the European regulation of AI (especially the proposed AI Act) with harmonized European Standards.

Standards play an important role in the planned AI Act. They serve to reliably implement the requirements of the AI Act and help make the development of AI systems more efficient and reliable. European Standards on transparency, logging, fairness, risk assessment and privacy protection, for example, are to be developed by the fall of 2024. The quality and speed of European standardization work depends on the substantive groundwork carried out at national level.

The central body for European AI standardization is the joint body **CEN/CENELEC JTC 21 "Artificial Intelligence" (CEN/CLC JTC 21)**[50], which was established by CEN and CENELEC on the basis of the recommendations of the "White Paper on Artificial Intelligence" [7] and the "German Standardization Roadmap Artificial Intelligence 1st Edition" [63] in spring 2021.

The CEN/CLC JTC 21 joint body is under German leadership and is supported by the secretariat held by Denmark. It is responsible for developing European Standards on artificial intelligence as well as advising other Technical Committees. Currently, the committee is dealing with the following topics, among others: Green and sustainable AI, Data Governance and Quality for AI, AI Systems risk catalogue and risk management, Overarching unified approaches on trustworthiness characteristics.

This Standardization Roadmap identifies concrete standardization needs for AI and thus provides significant support for both the national joint committee on AI (NA 043-01-42 GA) and European AI standardization.

Figure 13 shows the structure of the joint committee on AI (CEN/CLC JTC 21).

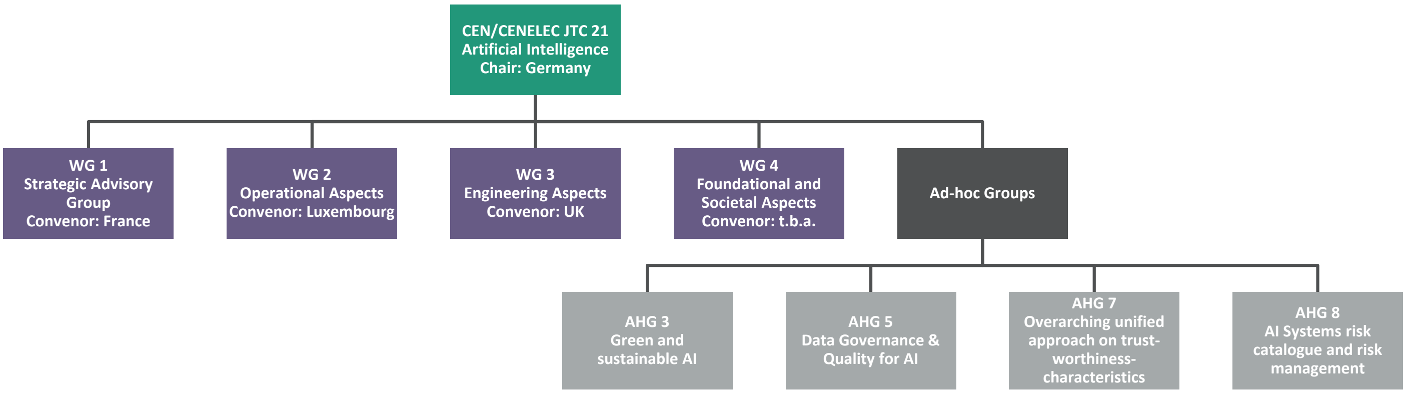


Figure 13: Structure of the European Joint Committee on AI (Source: DIN)

[49] See https://www.din.de/de/mitwirken/normenausschuesse/nia/nationale-gremien/wdc-grem:din21:54770248

[50] See https://www.cencenelec.eu/areas-of-work/cen-cenelec-topics/artificial-intelligence/

3.2.3 AI standardization at international level

At the international level, the joint committee **ISO/IEC JTC 1/SC 42 "Artificial Intelligence"**[51] was established in 2017 under US leadership.

This standardization body, set up by ISO and IEC, represents the central point of contact for AI standardization at international level. The currently 35 member countries of ISO/IEC JTC 1/SC 42 span all continents and are complemented by 15 "Observing Members". This global composition ensures an equally globally aligned work program, which currently includes standardization of AI fundamentals, data standards related to AI, big data and analytics, trustworthiness, policy implications of AI, and ethical and societal concerns. Thus, the body addresses the entire AI ecosystem and advises ISO and IEC committees on artificial intelligence.

Since its establishment, ISO/IEC JTC 1/SC 42 has already developed and published 14 International Standards. These include standards on big data[52] and the big data reference architecture[53], standards for evaluating the robustness of neural networks[54], for describing use cases[55], and for ethical and societal concerns[56].

The 25 standards projects currently underway address issues including data quality for analytics and machine learning, functional safety, and quality assessment guidelines and impact assessments for AI systems.

China, Ireland, Japan and Germany currently hold the secretariats of the "Working Groups" in ISO/IEC JTC 1/SC 42 and can thus actively shape the content-related work on AI standardization at international level as well.

Figure 14 shows the structure of ISO/IEC JTC 1/SC 42.

Further relevant international standards bodies include: ISO/IEC JTC 1/SC 27[57] "Information security, cybersecurity and privacy protection" is under German leadership and is responsible for the development of standards and specifications for the protection of information, and for information and communication technologies. This includes security and privacy aspects as well as cryptographic and other security mechanisms.

ISO/IEC JTC 1/SC 41[58] "Internet of things and digital twin" develops international standards and specifications on topics such as the Internet of Things, digital twins and related technologies.

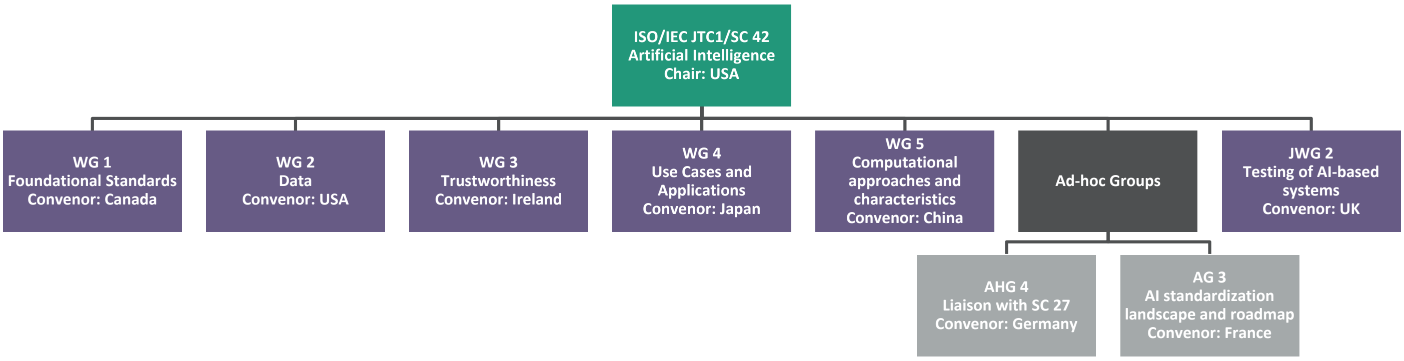


Figure 14: Structure of the international Joint Committee on AI (Source: DIN)

[51] See https://www.iso.org/committee/6794475.html

[52] See ISO/IEC 20546 [443]

[53] See ISO/IEC TR 20547 (series) [438], [439], [440], [441], [442]

[54] See ISO/IEC TR 24029-1 [91]

[55] See ISO/IEC TR 24030 [293]

[56] See ISO/IEC TR 24368:2022 [15]

[57] See www.iso.org/committee/45306.html

[58] See www.iso.org/committee/6483279.html

In addition to traditional, fully consensus-based standardization, specifications and recommendations on AI are also issued by some professional associations and consortia. Extensive consortial work on AI standardization emerges from various forums and consortia such as the IETF[59], IEEE[60], CSA Group[61], OGC[62], OMG[63], or W3C[64], for example, and complements standardization in sometimes very specialized subject areas.

3.3 Research and implementation projects on AI

Germany is one of the world leaders in research in the field of artificial intelligence. However, in order to sustainably leverage the potential of artificial intelligence and exploit it commercially, the innovative research results must also be transferred into practice. Being recognized and trusted strategic tools, standards and specifications can help provide rapid access to the market for scientific results.

At both national and European level, the standardization institutes are involved in AI research projects in various roles, thus supporting the identification of essential standardization potentials, the development of standardization strategies and the initiation of standardization activities.

A selection of AI research projects in which standardization is a core element is presented below.

3.3.1 AI research projects[65]

VIKING

The VIKING[66] project (Trusted Artificial Intelligence for Police Applications) started in January 2022 and is funded by the German Federal Ministry of Education and Research (BMBF). The goal is to develop a catalogue for compliance with acceptable ethical and high legal requirements for AI procedures in everyday policing, as well as operationalization of the underlying principles of trustworthy AI in light of new EU law and other legal and ethical requirements. With the interweaving of different activities in the project, the foundations for trustworthy AI are being created. Users are accompanying the work and evaluating the demonstrators and requirement catalogues with regard to functionality and practicality for everyday police work, which ensures that the research work is closely linked to the actual relevance for the interested parties. Where successful, the results from VIKING can advance to best practices for the police use of AI procedures in the future, strengthen law and security in Europe and significantly shape the rapidly growing national and international markets of this segment through “Technology Made in Germany”.

STAFFEL

Launched in December 2021, the STAFFEL[67] project was created by the German Federal Ministry of Transport and Digital Infrastructure (BMVI) and is dedicated to providing an AI-powered internet platform to enable secure, data-based and cross-shipper “staggered traffic”. To achieve this goal, the platform and the anti-theft system will be prototypically developed after a detailed requirements analysis and validated in two field tests. Initially, regional transport companies will be networked via a driving time marketplace. After that, alternating stations will be established along a main traffic route and relay traffic will be tested in practice. The aim is to identify effects, potentials and challenges for truck freight transport and to prepare for Europe-wide implementation. Standardization will also play an important role in this process.

BIG PICTURE

The Big Picture project[68] is funded by the European Union and since 2021 has aimed to enable the rapid development of AI in pathology by creating the first European ethical and quality-controlled platform in compliance with the GDPR (General Data Protection Regulation), where both big data and AI algorithms exist simultaneously. The BIGPICTURE platform is being developed in a sustainable and inclusive way by connecting communities of pathologists, researchers, AI developers, patients, and industry. Through the creation of a common infrastructure (hardware and software), millions of

[59] Internet Engineering Task Force, see: www.ietf.org/

[60] Institute of Electrical and Electronics Engineers, see: www.ieee.org/

[61] See: www.csagroup.org/

[62] Open Geospatial Consortium, see: www.ogc.org/

[63] Object Management Group, see: www.omg.org/

[64] World Wide Web Consortium, see: www.w3.org/

[65] The presentation makes no claim to completeness.

[66] https://www.din.de/de/forschung-und-innovation/partner-in-forschungsprojekten/ki/viking-872288

[67] https://www.din.de/de/forschung-und-innovation/partner-in-forschungsprojekten/ki/staffel-860360

[68] https://www.din.de/de/service-fuer-anwender/normungsportale/gesundheit/forschung-innovation-standards/aktuelle-forschungsprojekte/bigpicture-791128

images will be stored, shared and processed. Legal and ethical frameworks and functionalities are being established to ensure appropriate use and processing of data for diagnostic and research purposes, while fully respecting patient privacy and data confidentiality.

IMPULSE

The IMPULSE[69] (Identity Management in Public Services) research project is funded by the European Union under the Horizon 2020 programme and since 2021 has focused in particular on two of the most promising and disruptive technologies of our time: artificial intelligence (AI) and blockchain, and their contributions and implications for electronic identity management (eID) in public services. Two main deliverables will be produced: a holistic AI and blockchain technology supporting, GDPR-compliant eID, and actionable roadmaps and recommendations for the adoption, scale-up and sustainability of such advanced eID technologies by public services and for policymakers.

KIOptiPack

The KIOptiPack[70] project (holistic AI-based optimization of plastic packaging with recycled content) intends to develop and validate practical AI-supported tools for the successful and quality-oriented production of plastic packaging with a high recyclate content. The AI and data infrastructure will build on the concepts and systems developed in the Gaia-X initiative and enable distributed AI application and sovereign data sharing. AI-powered agile analytics tools will be used to support material qualification and increase quality, robustness and productivity in the production of packaging materials containing recyclates. Sustainability assessment and the further development of circular economy business models are pursued as an integral component of the project. In addition, an innovation laboratory involving all relevant stakeholder groups is to be established for the collaborative development of innovative solutions based on real consumer needs, and the necessary specifications and interests of the actors along the value chain.

ZVKI

The ZVKI[71] (Centre for Trustworthy Artificial Intelligence) project was launched in 2021 by the German Federal Ministry for the Environment, Nature Conservation, Nuclear Safety and Consumer Protection (BMUV) and aims to promote transparency and trust in AI applications among consumers through the interaction of politics, business, science and society. The activities of the ZVKI focus on education and the dissemination of information for consumers, as well as the scientific monitoring of AI applications with regard to their negative effects on the protection of humans. In addition, tools for evaluating AI systems and requirements for their certification are being developed to provide a basis for human confidence in AI developments and deployment. The goal is to bring together as many stakeholders and ideas as possible to jointly shape trustworthy AI.

KIDD

Since 2020, the focus of the BMAS-funded project KIDD[72] (AI in the Service of Diversity) has been to enable companies to introduce human-centric digital applications into their operations. Here, an innovative process (KIDD process), transferable to other companies and organizations, is being developed for the transparent, participatory and inclusive introduction of AI in companies. Furthermore, the how of digitalization will be discussed and tested in a fair, transparent and understandable way, and the concrete results will be made available to a broad public after completion in order to make digitalization in companies fair and transparent.

KIMEDS

The KIMEDS[73] (AI-assisted certification of medical software) project, funded by the BMBF and launched in 2022, aims to improve certification procedures for software-based medical technology. The project is researching AI systems to help speed up the approval process. In particular, the monitoring of product safety risks, from the development of medical software to certification and monitoring in operation, will be supported by AI systems. These certification processes often pose a major challenge, especially in medicine, when it comes to complying with existing regulations. The project aims to answer the question of how an AI system can adequately support this process.

[69] https://www.din.de/de/forschung-und-innovation/partner-in-forschungsprojekten/ki/impulse-799412

[70] https://www.fona.de/de/massnahmen/foerdermassnahmen/ki-hub-kunststoffverpackungen.php

[71] https://www.zvki.de/

[72] https://kidd-prozess.de/

[73] https://tu-dresden.de/tu-dresden/newsportal/news/zertifizierung-medizinischer-software-mit-ki-grundlegend-verbessern

QI-Digital

The goal of the QI-Digital[74] project is to design a reliable quality infrastructure (QI). As a system in various institutions and processes, it contributes significantly to the safety of products and applications, the protection of health and the environment, and the functioning of trade in goods and services. The QI-Digital initiative, launched in 2021 and consisting of the partners BAM, DAkkS, DIN, DKE and PTB, is developing a set of fields of action together with network partners from science and industry as well as other QI actors and is working out practical solutions for concrete case studies of economically significant technologies and innovations. To this end, a comprehensive QI digital innovation ecosystem is being created to provide the foundation and framework for the development and establishment of practical solutions. Exemplary for the envisioned QI Digital innovation ecosystem, the QI Digital initiative is working on very specific projects. AI in medical technology, additive manufacturing and modern hydrogen applications are three innovation fields in which test field environments have been started. Quality and safety – and the resulting trust of all stakeholders – are crucial to the success of these future technologies.

3.3.2 Projects implementing the Standardization Roadmap AI

In addition to the classic research projects, the standardization institutes are working with various partners on projects which implement the recommendations for action from the Standardization Roadmap Artificial Intelligence. The implementation projects look at use cases that are typical of applications and relevant to industry, and identify requirements for the standardization of AI-specific applications. With the help of these projects, practical experience is to be gathered in the respective application context, concrete needs for standardization are to be derived, and findings on quality and conformity testing are to be obtained.

Among the implementation projects, the Coordination Group “AI Standardization and Conformity” designates “lighthouse projects” based on defined criteria. They are of particular importance in the implementation of the Standardization Roadmap AI, which is why they are gaining increased attention among standardization actors and are widely visible in industry, research and politics.

ZERTIFIZIERTE KI

The project ZERTIFIZIERTE KI[75] is a project implementing the 1st edition of the Standardization Roadmap AI. The goal of the project, which started in 2021, is to develop and standardize test criteria, methods, and tools for AI systems, thus enabling a comparable evaluation of AI systems. The verifiability of technically guaranteed properties is intended to increase the trust of users and consumers in AI technologies. In sector- and technology-related user groups, participants from business and industry as well as science will define concrete requirements, establish criteria and benchmarks for testing in practice, and verify them on the basis of use cases. A broad participatory process will be used to ensure that procedures evolve into generally accepted standards for AI systems and their verification, while taking into account legal, ethical, and philosophical considerations.

safetr.AIn

The safe.trAIn project[76] (Safe AI using the example of driverless regional trains) is the first official lighthouse project[77] of the Standardization Roadmap AI. It is funded by the BMWK and since 2022 has been pursuing the goal of linking AI processes with the requirements and approval processes in the rail environment in a practicable manner. The focus of the consortium, consisting of representatives from the rail industry, technology suppliers, research institutes and standardization and testing organizations, is on the development of standardized test methods and tools to ensure approval-relevant product safety for the widespread use of fully autonomous trains. In addition, the safety architecture is being given detail using the example of the driverless regional train, and a fully automated GoA4 system is being conceptually developed and validated for this use case in a virtual test field. Standards and specifications play a decisive role in accelerating time-to-market and the safe, robust, and trustworthy application of AI-based methods for driverless train travel.

[74] https://www.din.de/de/din-und-seine-partner/presse/mitteilungen/qi-digital-792188

[75] https://www.din.de/de/vertrauen-in-ki-staerken-mit-qualitaetskriterien-und-pruefverfahren--791046

[76] https://www.din.de/de/forschung-und-innovation/partner-in-forschungsprojekten/ki/safe-train-860442

[77] See Chapter 6.6.

AI readiness of standards

Since 2022, the AI readiness of standards[78] project, led by DIN, has been pursuing the particular goal of identifying and describing the content-related relevance of relevant standards to artificial intelligence. AI technologies are already being used in almost all disciplines – including those where standards are applied without being designed for them. In order to enable the progress of AI technologies in all disciplines, an analysis of the entire body of standards and, if necessary, an adaptation of the relevant standards is necessary. The project will develop a scalable methodology for analyzing the body of standards with regard to any points of contact with AI technologies in practice. To complement this, a software-based AI tool is being developed to assist in this analysis in the future to identify relevant standards. In addition, the development of machine executability of standards (SMART standards, see Chapter 5.3) is being supported by elaborating requirements of AI systems on the structure of standardization documents. More detailed information about the project is given in Chapter 5.1.

[78] https://www.din.de/de/forschung-und-innovation/themen/kuenstliche-intelligenz/projekte-zu-ki-und-normung/ki-tauglichkeit-von-normen/ki-tauglichkeit-von-normen-872324

4 Key topics

Artificial intelligence is a cross-sectional technology that is already being used in diverse areas today and is thus influencing almost the entire economy and society. The scope and complexity of this topic do not allow all areas to be considered within the scope of this Standardization Roadmap. Therefore, it has a targeted focus and is structured according to horizontal topics and addresses the relevant industrial sectors and application areas. Figure 15 shows the key topics of this Standardization Roadmap and gives the structure of this chapter.

New technological developments are raising questions about overarching issues, particularly in the application of AI.

The starting point here are **fundamental topics** such as terminologies (definitions), AI classifications, and ethics. They are the basis for all discussions about AI and thus represent one of the horizontal key topics (see Chapter 4.1).

The aspect of **security/safety** is becoming increasingly important in the context of AI – both in terms of protection against external attacks (security) and freedom from errors or operational reliability (safety). Only a deeper consideration of the security/safety of AI-based technologies and applications can enable their comprehensive use in industry and society (see Chapter 4.2).

Another key requirement for the widespread use of AI systems is **testing and certification**. They can be instrumental in building trust in AI systems and creating acceptance. Chapter 4.3 provides insights into the current state of discussion on assessing the quality of AI applications.

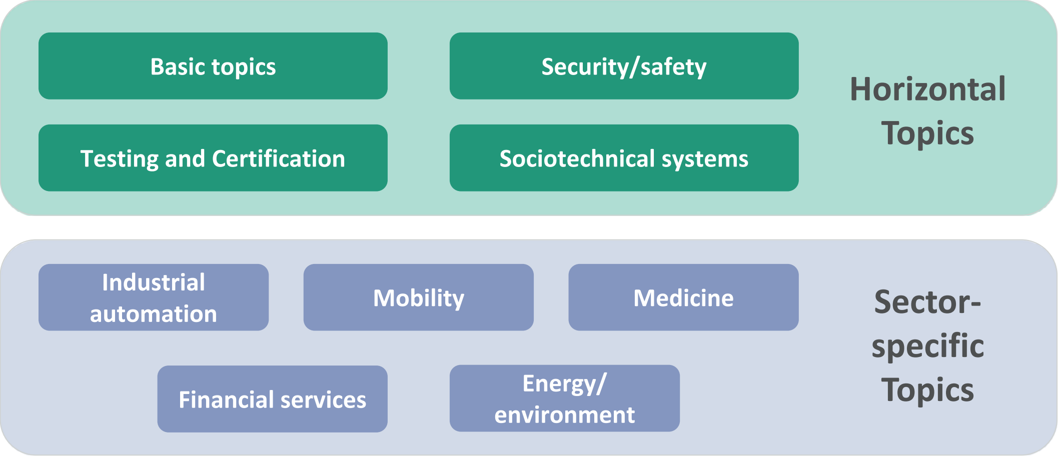
The last horizontal topic to be considered is **sociotechnical systems**. The focus here is particularly on the human-technology interface. Important issues include the integration of AI technology into societal subsystems, human-technology interaction, and organizational development (see Chapter 4.4).

The industrial sectors and application areas of AI are extremely diverse. In all areas, AI technologies offer great potential.

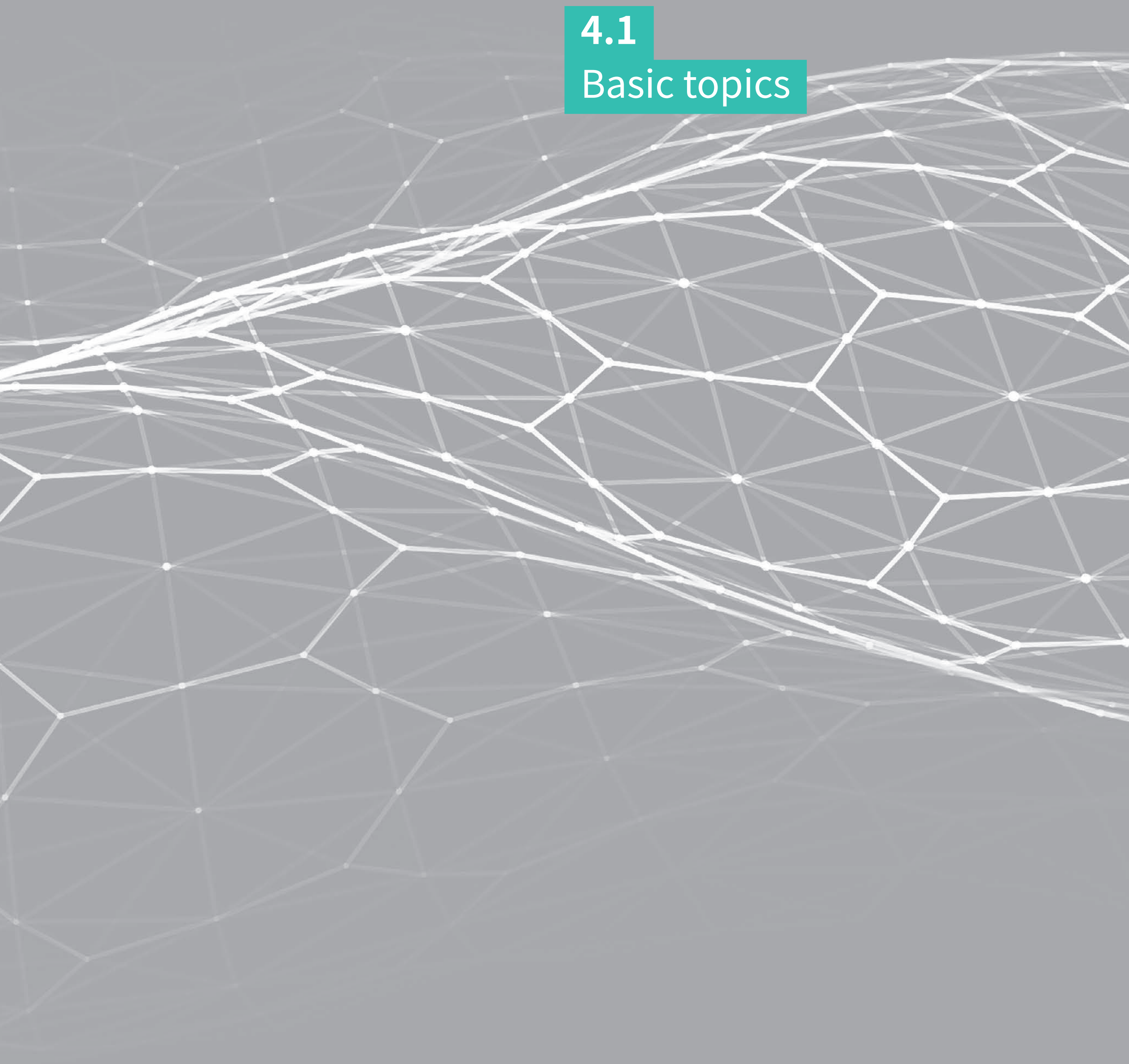
In addition to the overarching topics, this edition of the Standardization Roadmap AI focuses on the five application areas of **industrial automation, mobility, medicine, financial services, and energy/environment,** which covers as broad and diverse a spectrum of applications as possible (see Chapter 4.5 to Chapter 4.9).

In the following, the initial situation, requirements and challenges, as well as concrete standardization needs, are elaborated for the nine key topics of the Roadmap.

Figure 15: Overview of key topics (Source: DIN)



4.1 Basic topics



AI is a cross-cutting topic that touches on many disciplines, some of which are considered in the Chapters 4.2 to 4.9. For a basic understanding, a definition of terms for the present document has already been given in Chapter 1.5; in addition to this, superordinate topics are dealt with in the following chapter.

4.1.1 Status quo

There are already numerous activities in the standardization environment in the field of AI fundamentals. The most important committees in this regard have already been presented in Chapter 3.2, with particular emphasis on the work of ISO/IEC JTC 1/SC 42 [14], which is mirrored by the DIN/DKE Joint committee NA 043-01-42 GA Artificial Intelligence. A selection of the most significant projects is listed below. Furthermore, Chapter 4.1.1.1 presents the current status of the classification of AI.

Title	Contents	Status
ISO/IEC 22989:2022, Artificial intelligence – Concepts and terminology [16]	Concepts and terminology for artificial intelligence	Published
ISO/IEC 23053:2022, Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML) [24]	Terminological framework for machine learning	Published

On the topic of AI systems management, the following works are to be mentioned:

Title	Contents	Status
ISO/IEC 23894:2022, Information Technology – Artificial Intelligence – Guidance on risk management [25]	Guidelines for the risk management of the development and use of AI systems. This standard is also being developed under the direction of a German editor.	In development, publication due end 2022
ISO/IEC 38507:2022, Information technology – Governance of IT – Governance implications of the use of artificial intelligence by organizations [26]	Organizational governance in connection with AI	Published
ISO/IEC 42001, Information Technology – Artificial Intelligence – Management System [27]	Certifiable management standard for AI that contains requirements and organizations for the responsible development and use of AI systems.	In development, publication due mid 2023
ISO/IEC 42005, Information Technology – Artificial Intelligence – AI System impact assessment [432]	Impact assessment for the use of AI systems	Initiated, publication due 2025
Information technology – Artificial intelligence – Requirements for bodies providing audit and certification of artificial intelligence management systems	Requirements and certification bodies	Initiated, publication due 2024

The topic of ethics (to the extent not already addressed by the above documents) is addressed in the following:

Title	Contents	Status
ISO/IEC TR 24028:2020, Information technology – Artificial intelligence – Overview of trustworthiness in artificial intelligence [28]	Overview of the trustworthiness of AI systems	Published
ISO/IEC TR:24368:2022, Information technology – Artificial intelligence – Overview of ethical and societal concerns [15]	Overview of ethical topics relating to the work programme of SC 42	In development, publication due mid 2023

The development of AI systems and system-specific aspects of their use and evaluation are the subject of the following standardization projects:

Title	Contents	Status
ISO/IEC TS 4213 Information technology – Artificial intelligence – Assessment of machine learning classification performance [29]	Metrics for machine learning performance	TBD
ISO/IEC 5338 Information technology – Artificial intelligence – AI system life cycle processes [30]	Life cycle processes, based on the life cycle model in ISO/IEC 22989:2022 [16]	TBD
ISO/IEC 5339 Information Technology – Artificial Intelligence – Guidelines for AI applications [31]	Recommendation for the use of AI systems (also addresses ethical aspects)	TBD
ISO/IEC 5392 Information technology – Artificial intelligence – Reference architecture of knowledge engineering [32]	Reference architecture for symbolic AI systems	TBD
ISO/IEC TR 5469 Artificial intelligence – Functional safety and AI systems [33]	Overview of the functional safety of AI systems	TBD
ISO/IEC TS 5471 Artificial intelligence – Quality evaluation guidelines for AI systems [34]	Recommendations for the quality evaluation of AI systems, based on the SQuaRE model	TBD
ISO/IEC 25059:2022 Software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – Quality model for AI systems [35]	Requirements for the quality evaluation of AI systems, based on the SQuaRE model	TBD
ISO/IEC TS 6254 Information technology – Artificial intelligence – Objectives and approaches for explainability of ML models and AI systems [36]	Overview and recommendations for dealing with the explainability of AI systems	TBD

Title	Contents	Status
ISO/IEC TS 8200 Information technology – Artificial intelligence – Controllability of automated artificial intelligence systems [37]	Overview and recommendations for dealing with the controllability of AI systems	TBD
ISO/IEC TS 12791 Information technology – Artificial intelligence – Treatment of unwanted bias in classification and regression machine learning tasks [38]	Recommendations on how to avoid unwanted bias for classification and regression	TBD
ISO/IEC 12792 [238] Information technology – Artificial intelligence – Transparency taxonomy of AI systems	Recommendations on the documentation of transparency requirements for AI systems	TBD

Data quality management is dealt with in the following standards:

Title	Contents	Status
ISO/IEC 5259-1 Artificial intelligence – Data quality for analytics and machine learning (ML) – Part 1: Overview, terminology, and examples [40]	Part of a standards series on data quality management: Part 1 describes terminology and concepts which will be used in further parts in the series.	TBD
ISO/IEC 5259-2 Artificial intelligence – Data quality for analytics and machine learning (ML) – Part 2: Data quality measures [41]	Part of a standards series on data quality management: Part 2 deals with quality measures.	TBD
ISO/IEC 5259-3 Artificial intelligence – Data quality for analytics and machine learning (ML) – Part 3: Data quality management requirements and guidelines [42]	Part of a standards series on data quality management: Part 3 addresses requirements.	TBD
ISO/IEC 5259-4 Artificial intelligence – Data quality for analytics and machine learning (ML) – Part 4: Data quality process framework [43]	Part of a standards series on data quality management: Part 4 describes processes that can be implemented to fulfil the requirements in Part 3.	TBD
ISO/IEC 5259-5 Artificial intelligence – Data quality for analytics and machine learning (ML) – Part 5: Data quality governance [44]	Part of a standards series on data quality management: Part 5 deals with the governance of data.	TBD
ISO/IEC 8183 Information technology – Artificial intelligence – Data life cycle framework [45]	Part of a standards series on data quality management (note the different standard number): This part deals with the data life cycle to supplement Part 1 of the ISO/IEC 5259 series [39]	TBD

4.1.1.1 AI classification

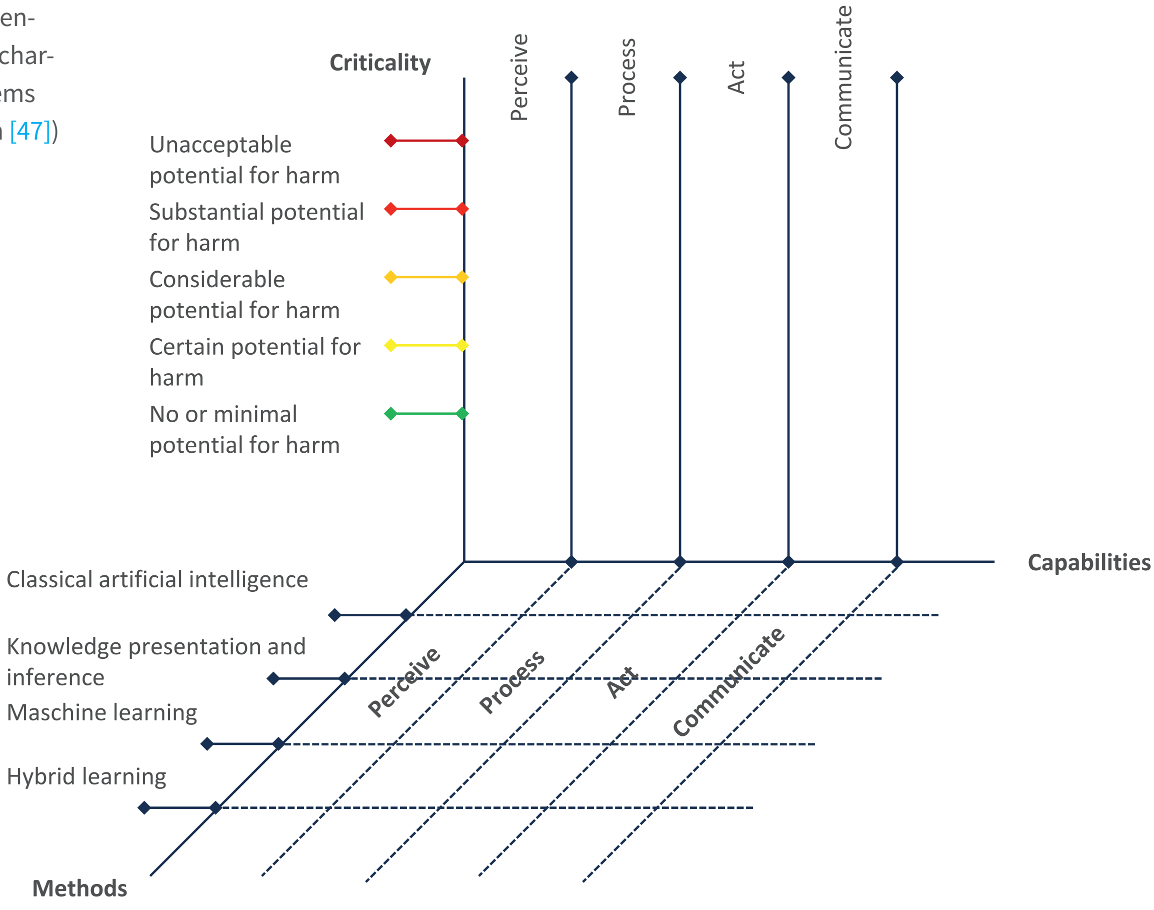
According to the position paper “A definition of AI: Main capabilities and scientific disciplines” of the AI HLEG [46], a distinction is made between the methods and capabilities of AI. In both cases, the following classifications are based on a recent review paper [47], which originated around the time of the first version of the Standardization Roadmap AI and which reflects the current state of the art. The section “Classification of AI Methods” describes which AI methods are used to realize specific AI capabilities. The section “Classification of AI capabilities” describes basic capabilities of AI systems. In combination with a criticality assessment or risk assessment, this classification scheme enables a holistic characterization of an AI system (Figure 16). In order to also adequately reflect the actual state of the current industrial AI markets, an additional classification of AI applications is made based on AI methods and AI capabilities. Further information and examples can be found in the recently published Beuth Pocket title [48].

Classification of AI methods

Today’s AI is based on a variety of different methods. Based on historical developments, a rough distinction is often made between symbolic and subsymbolic AI methods. Both paradigms form the basis of a large number of AI applications in practice ([49], 79-88). As regards symbolic methods, techniques of knowledge representation and logical reasoning are in the foreground, while sub-symbolic methods are represented mainly by machine learning techniques and neural networks. However, this traditional distinction is not comprehensive. It neglects classical methods of artificial intelligence such as problem solving, optimization, planning and decision-making. Moreover, developments in recent decades have further blurred traditional boundaries, and combined or hybrid approaches are increasingly coming to the fore, e.g., the entire field of hybrid learning ([50], p. 77; [49]).

Table 2 and Table 3 provide an overview of these methods at three levels of granularity (fields of AI, disciplines of AI, and sub-disciplines) and name well-known examples of each

Figure 16: Three-dimensional scheme for the characterization of AI systems (Source: adapted from [47])


Criticality
Perceive
Process
Act
Communicate
Unacceptable potential for harm
Substantial potential for harm
Considerable potential for harm
Certain potential for harm
No or minimal potential for harm
Capabilities
Classical artificial intelligence
Knowledge presentation and inference
Maschine learning
Hybrid learning
Perceive
Process
Act
Communicate
Methods

technique. This scheme is a preliminary snapshot that may be augmented by emerging AI methods in the future. Also, it is often not possible to draw an incontrovertible distinction between categories, as some methods may belong to more than one category. In such cases, methods were assigned according to Russell and Norvig [51] or according to the category for which they were originally proposed.

(A) Classical artificial intelligence

Historically, approaches to problem solving, optimization, planning, and decision-making were among the earliest AI methods to be developed. Problem solving describes goal-oriented search strategies and intelligent agents that solve problems by formulating a goal and, with a defined problem as input, by searching for the correct sequence of actions to execute the solution and achieve the goal. In competitive multi-agent environments where the goals are in conflict, adversarial and constraint-based searches are used to solve complex problems.

Unlike problem-solving methods that systematically explore search spaces, optimization algorithms do not care about the path to the goal, but focus on the optimal solution. They can be divided into deterministic approaches such as simplex methods, network algorithms, decision trees (including branch-and-bound methods), and classical gradient descent methods, and non-deterministic approaches. Examples of non-deterministic optimization methods include genetic algorithms, swarm intelligence, and simulated cooling.

Planning methods can be autonomous or semi-autonomous techniques, such as steady-state searching, planning graphs, hierarchical planning, non-deterministic planning, time and resource planning, and plan generation. In contrast to planning, plan recognition models or methods such as deductive and synthesis plan recognition, library-based plan recognition, and planning by abductive inference must represent actual events or actions that occurred and propose hypothetical explanations. Planning methods play a role in robotics, dialogue systems, and human-machine interaction.

Decision-making or decision analysis is a technical discipline concerned with the pragmatic application of decision theory to particular problems ([52], 247-302). There are various approaches to decision-making such as process models, information value, decision networks, expert systems, sequential decision-making, and iterative models.

(B) Symbolic artificial intelligence

Symbolic AI methods are characterized by a deductive approach, i.e., the algorithmic application of logical rules or relationships to individual cases. Core concepts of symbolic AI are, first, techniques for representing knowledge and, second, methods for applying that knowledge to a given input. Knowledge can be represented as either certain or uncertain knowledge. With the help of argumentation chains, conclusions can be drawn from this knowledge.

Formal knowledge representation includes concepts such as ontologies, semantic networks, knowledge graphs, and knowledge maps that summarize and systematize information into structures, syntaxes, semantics, and semiotics. The focus of standardized description languages such as the Resource Description Framework (RDF) and the W3C Web Ontology Language (OWL) is on creating unique specifications for objects, features, and general concepts through logical relationships. Using these and other semantic web standards and technologies, logically related data can be shared across domains of different applications, facilitating semantic interoperability. In general, the basic concepts of ontology engineering are based on taxonomy, calculus, deduction, abduction, and the processing and modelling of ontologies. In addition, logical relationships and abstractions of domains can be established through knowledge graphs, semantic networks, and knowledge mapping. In the case of graph-based abstraction of knowledge, graph traversal algorithms provide common solutions to problems of searching, verifying, and updating vertices. Furthermore, logically related data can be modelled by propositional or predicate logics, higher-order logics, non-monotonic, temporal, and modal logics.

For certain knowledge, the application of formal knowledge is often operationalized using the classical methods of logical inference. In particular, satisfiability and other techniques of formal verification can be applied here. Probabilistic approaches are widely used for inference based on uncertain knowledge, but non-probabilistic approaches have also been proposed. In probabilistic inference, information can be inferred by sampling a knowledge base and processed through relational probabilistic models or the concept of Bayesian inference. With respect to uncertain knowledge, Bayes' rule dominates the AI field in quantifying uncertainty. Non-probabilistic inference can be applied to ambiguous information in the case of vagueness with the consideration of evidence. In these situations, a truth management system and inferences with standard information can be used for qualitative approaches. In addition, methods of non-probabilistic inference can be implemented using rule-based approaches or fuzzy sets. Another common approach for non-probabilistic inference is the inference with belief function, where all available evidence is combined to calculate a degree of belief. Other approaches to uncertain inference include spatial inference, case-based inference, qualitative inference, and psychological inference.

(C) Machine learning

In contrast to symbolic AI, subsymbolic AI methods are characterized by an inductive approach, i.e., by the algorithmic derivation of general rules or relations from individual cases. To this end, two broad approaches to machine learning are typically distinguished: supervised learning with given target parameters and unsupervised learning where these parameters are not given. In addition to these two main approaches, alternative learning paradigms such as semi- supervised, reinforcement, or counterfactual learning have also become established.

Supervised learning techniques are typically used to perform regression or classification tasks. Practical applications of supervised learning have long been dominated by discriminative methods such as logistic regression, decision trees, or neural networks. Neural networks are considered to be particularly flexible, since they can theoretically learn any mathematical function without any prior knowledge. Support vector machines are also successfully used in many applications, despite the necessary, but not easily determined, kernel function.

Some supervised learning algorithms such as Naive Bayes or Hidden Markov models produce an estimated probability distribution for input and output variables. Although these are widely counted among supervised learning techniques, they can be more accurately referred to as generative learning techniques. More recent examples of generative methods include techniques such as generative adversarial networks, but these follow a counterfactual learning paradigm. This paradigm originates from the application domain of image processing and represents an evolution of classical supervised learning, which relies on two machine learning methods operating against each other to recreate the features of a given dataset.

Unsupervised learning methods, on the other hand, are usually used for clustering or dimension reduction tasks. One of the oldest and best known algorithms is the clustering method k-means. In addition to other statistically motivated methods such as hierarchical clustering, biologically inspired algorithms such as Kohonen's self-organizing map and Grossberg's adaptive resonance theory have also been proposed. There are also unsupervised methods for regression tasks, which are mainly used for dimension reduction.

Not all learning algorithms can be clearly classified as supervised or unsupervised. For example, a multilayer perceptron, i.e., a supervised method, can be used to map a given dataset to itself. If one subsequently removes the output layer of such an autoencoder, what remains is a subnetwork that performs a dimensional reduction of the dataset according to the number of hidden neurons. Such an application of supervised learning methods for the targeted generation of specific representations in individual layers of neurons is, for example, an important basis of deep learning. Another example of intermediate forms of machine learning are algorithms of semi-supervised learning, where supervised and unsupervised learning are combined and thus a predefined target value is only required for a part of the data used. This not only enables the analysis of incomplete datasets, but in some cases even achieves better results than classical supervised learning methods. However, for semi-supervised learning algorithms, assumptions about distribution densities must be made in advance. If the assumptions are unfavourable, the results can be significantly worse than with a supervised learning procedure.

A related but alternative learning paradigm to supervised learning is what is known as reinforcement learning. This requires feedback for predictions made, but not the exact target value, for context learning. Analogous to learning in the form of classical conditioning (i.e., via reward or punishment), this method only takes into account whether the intended learning outcome was achieved or not. Such learning with feedback but without a fixed target value has proven to be very useful, especially in the application areas of robotics and adaptive control.

(D) Hybrid learning

Hybrid learning methods are characterized by the combination of concepts from the previously presented methods, e.g., training neural networks by applying genetic algorithms to adjust the network weights and, if necessary, the network architecture. This combined approach has been proposed for applications as diverse as financial applications, oceanographic predictions, ambulance visit predictions, and tea plant classification.

Due to the scientific creativity in this field, it is difficult to impossible to give a comprehensive overview of all methods of hybrid learning. However, much of this field focuses on combining symbolic and subsymbolic AI to create both inductive and deductive systems. Recent research activities address, for example, the combination of machine learning and knowledge engineering [53]. A prominent subfield is hybrid neural systems, which can be further divided into unified neural architectures, transformational architectures, and hybrid modular architectures ([54], 62-93). In contrast to classical subsymbolic methods, such methods allow either the extraction of rules or use an additional form of knowledge representation. However, unlike classical symbolic methods, such knowledge representations are often algorithmically created or modified based on given data.

In a broader sense, hybrid learning can also be described as learning with knowledge. To this end, other approaches have been proposed in recent decades, such as learning by logic and deduction, inductive logic programming, explainable AI, and relevance-based learning. More recent approaches to hybrid AI include what is known as conversational learning or active dialogue learning, which aims to improve machine learning performance by incorporating human knowledge gathered through dialogue.

Table 2: Classification of AI methods

Field	Discipline	Subdiscipline	Examples
CLASSICAL ARTIFICIAL INTELLIGENCE	Problem solving	Agents & search strategies	Uninformed & informed search strategies
			Inverse searching (game theory)
			Searching under boundary conditions
	Optimization	Deterministic	Simplex methods
			Network algorithms
			Decision trees (e.g. branch & bound)
			Gradient descent methods
		Non-deterministic	Evolutionary algorithms
			Genetic algorithms/programming
			Swarm intelligence
			Simulated cooling
	Planning & plan recognition	Autonomous & semi-autonomous planning	Steady-state searching
			Planning graphs
			Hierarchical planning
			Non-deterministic planning
			Time & resource planning
			Plan generation

Field	Discipline	Subdiscipline	Examples
		Plan recognition	Abductive plan recognition
			Deductive plan recognition
			Library-based plan recognition
			Synthesis plan recognition
	Decision-making	Approaches to decision-making	Process models
			Information value
			Decision networks
			Expert systems
			Sequential decision-making
			Iteration models
SYMBOLICAL ARTIFICIAL INTELLIGENCE	Knowledge representation	Ontologies	RDF, RDFS and OWL
			Taxonomies
			Interpretation
			Calculus
			Deduction
			Abduction
			Ontology mapping
		Knowledge graphs & semantic networks	Knowledge graphs and networks
			Existential graphs
			Graph traversal algorithms
			Mapping
			Semantic web
		Modelling by means of formal logic	Propositional logic & predicate logic
			Higher-order logics
			Non-monotonic logics
			Temporal & modal logics
		Quantification of uncertainty & representation of uncertain knowledge	Bayes' rule
			Bayesian networks

Field	Discipline	Subdiscipline	Examples
	Logical inference	Formal verification	Resolution & connectivity verification
			SAT & SMT (satisfiability modulo theories) solvers
			Model checking
		Interactive verification	Tactical theorem verification
	Probabilistic inference	Bayesian inference	Precise inference
			Approximate inference
			Markov chains
		Relational probabilistic models	Relational probabilistic models in closed & open universes
		Probabilistic inference with time & uncertainty	Hidden Markov model
			Kalman filter
			Dynamic Bayesian networks
	Non-probabilistic inference	Qualitative approaches	Inference with standard information
			Truth management systems
		Rule-based approaches	Rule-based inference with certainty
		Inference with uncertainty	Fuzzy sets & logic
		Inference with belief function	Dempster-Shafer theory
	Further approaches for uncertain inference		Spatial inference
			Case-based inference
			Qualitative physics
			Psychological inference

Field	Discipline	Subdiscipline	Examples
MACHINE LEARNING	Supervised learning	Neural networks	Multi-layer perceptron
			Learning vector quantization (LVQ)
			Radial basis networks (RBF)
			Adaptive resonance theory (ART)
			Convolutional neural networks (CNN)
			Recurrent neural networks (RNN)
			Time delay networks (TDNN)
			Long-short term memory (LSTM)
			Hopfield networks
			Boltzmann machines
		Statistical learning	Decision trees
			Random forests
			Support vector machine (SVM)
		Probabilistic methods	Naive Bayes
			Fuzzy classifiers
	Unsupervised learning	Clustering	k-means
			Hierarchical clustering
			DBSCAN
			Fuzzy clustering
			Self-organizing maps
		Dimension reduction	Autoencoders
			Primary component analysis
		Probabilistic methods	Fuzzy c-means

Field	Discipline	Subdiscipline	Examples
	Semi-supervised learning	Statistical methods	Expectation maximization with generative models
			Transductive support vector machines
		Modified learning concepts	Self-learning
			Collaborative learning
		Graph-based learning	Graph-based methods
	Reinforcement learning	Temporal difference learning	Q learning
			State–action–reward–state–action (SARSA)
		Monte Carlo methods	Markov chains Monte Carlo
		Adaptive dynamic programming	Active & passive adaptive dynamic programming
	Gegenläufiges Lernen	Generative methods	Generative adversarial networks (GAN)
			Bayesian adversarial networks
			Inverse autoencoders
		One-shot learning	Siamese neural networks
HYBRID LEARNING	Hybrid neural systems	Neural unit architectures	Constructivitistic Machine learning
		Transformation architectures	Rule extraction for neural networks
		Hybrid module architectures	Neural-fuzzy expert systems
	Learning with knowledge	Learning with logic & inference	Current best learning
		Inductive (logic) programming	Sequential covering algorithm
			Constructive induction algorithm
		Explainable artificial intelligence	Local interpretable model-agnostic explanations (LIME)
		Relevance-based learning	
	Conversational learning	Learning via active dialogue	Supervised conversational learning
			Reinforcement conversational learning

Classification of AI capabilities

The main inspiration for the establishment of the scientific discipline of artificial intelligence are the cognitive abilities of humans [51]. Psychology and cognitive science usually focus on partial aspects and less on a holistic view of these abilities. In educational contexts, on the other hand, human capabilities have been defined and assessed on the basis of learning goals since the middle of the last century. Learning goal taxonomies such as Benjamin S. Bloom's classify human abilities and form the basis of European education systems [55]. Bloom first distinguishes cognitive, affective, and psychomotor abilities [56], which in turn can be subdivided into more specific abilities. Based on Bloom's ideas, separate taxonomies have been postulated for affective abilities, i.e., behaviour predominantly governed by brief, impulsive emotional responses rather than cognitive processes, for psychomotor abilities, i.e., muscle control and coordination, and for cognitive abilities. This tripartite division is reflected in the following by the three general abilities of perceiving, processing, and acting (cf. Table 2).

Measured against such taxonomies, contemporary AI technology realizes only a subset of human cognitive capabilities. At the same time, many existing AI systems and applications allow for additional functionalities, such as enabling sensing beyond human senses or interaction with the environment (e.g., non-human communication). Based on these observations, both existing and potentially achievable AI capabilities can be broadly categorized as perceiving, processing, acting, and communicating. While this four-way division of capabilities takes into account common findings from psychological and educational research, it is primarily intended to provide a structure for capabilities that can be implemented by AI systems today. Table 3 gives an overview of the proposed classification of AI capabilities.

(A) Perceive

The term "perception" classically refers to abilities enabled by the human sensory organs. The ancient Greek philosophers, for example, distinguished the five senses of sight, hearing, smell, taste and touch. Modern science, on the other hand, distinguishes between sensory organs which transmit sensory stimuli and function as a type of perceptual precursor, and sensory modalities which essentially describe the output of the sensory organs for subsequent cognitive processing. With regard to AI systems, it should be noted that the variety of physical quantities that can be perceived by specialized technical sensors exceeds the number of stimuli that can be perceived directly by humans. Technical sensors exist today for a wide range of acoustic, biological, chemical, electrical, magnetic, optical, mechanical, radiant and thermal stimuli.

To describe human perception, many scientists today focus on sensory modalities. The term modality is often used to describe the encoding or "mode of representation" that results from the transduction of a sensory input. With the increasing popularity of a modality-based perspective on perception, the classic five-sense scheme has been questioned, and several alternative schemes have been proposed in recent decades. While the number of senses in such alternative classification schemes varies (between 8 and 17), there seems to be a consensus that in addition to the classical five senses, which are directly focused on external impressions, internal senses such as body perception and balance also contribute to human cognition by, for example, detecting disturbances and abnormalities ([57], 353-370). The sense of touch is also often considered a multifaceted sense that includes the perception of temperature, pressure, and pain [58].

Accordingly, there is a proposal to align AI capabilities in perception with human sensory modalities. For example, it should be noted that AI research has recently made significant progress in the ability to convert images, auditory, and haptic signals into processable information. AI applications for olfactory and taste perception, on the other hand, have been studied but are still relatively rare in practice.

(B) Process

The ability to process information is an essential prerequisite for intelligent behaviour. To describe this ability in more detail, it is convenient to use an existing taxonomy of cognitive learning goals such as Bloom's revised taxonomy ([59], 212-218) as a classification scheme. This taxonomy of learning goals, updated from Bloom's original taxonomy, distinguishes human abilities in a primary dimension according to the four domains of factual, conceptual, procedural, and metacognitive cognition.

Factual cognition, as the least complex domain in this schema, includes abilities to process and understand terminologies and knowledge of specific details and elements. Conceptual cognition includes the ability to process and understand knowledge about classifications and categories, principles and generalizations, and knowledge about theories, models, and structures. Procedural cognition refers to the processing of knowledge about subject-specific skills and algorithms, about subject-specific techniques and methods, and about criteria for determining the right time to apply appropriate methods. Metacognitive cognition includes abilities to process and understand strategic knowledge, knowledge about cognitive tasks (including contextual and conditional knowledge), and self-knowledge.

On a second level, these areas are further differentiated by six cognitive stages. For example, conceptual knowledge separates the basic skills of recognizing or classifying from the intermediate-complex skills of providing or discriminating information and the advanced skills of determining or assembling information. Using both cognitive dimensions, Bloom's revised taxonomy allows the discrimination of up to 24 human cognitive abilities. In terms of currently realizable AI capabilities, this allows, in particular, the mapping of the ability to reproduce knowledge, make decisions or make predictions. These capabilities are at the core of many advanced AI systems and are often used in combination with capabilities for perception, action, or communication.

(C) Act

For humans, the ability to act is a fundamental ability. In a more general sense, however, an action can be related to both a human and a non-human actor. It can be described as something an actor does that was "intentional under a certain description" [60]. Furthermore, a distinction can be made between physical and non-physical action. This is often realized by combining mechatronic and software components in robots or software robots. In particular, the field of robotics describes mechanically or physically performed activities such as robot perception, motion planning, sensors and manipulators, kinematics and dynamics, as well as the field of human-robot interaction, since this form of interaction focuses on physical human-machine interaction. These abilities are roughly based on the human abilities of controlling and coordinating muscles. The methods used for software agents depend on the particular goal or task of the agent itself. Such autonomous agents are indispensable, for example, in the field of process automation.

(D) Communicate

Although communication is an ubiquitous and well-researched human capacity, communication researchers have traditionally found it difficult to agree on a common definition or taxonomy. One of the simplest technologically motivated definitions sees communication as the transmission of information between specific subjects ([61], 379-423). In a complementary approach, on the other hand, communication is defined by the ability to communicate: This is then described as the ability to process an expression, to understand it, and to distinguish between expression and information, and thus consequently to distinguish between "the informational value of its content" and "the reasons for which the content was expressed" ([62], 251-259). In this sense, the ability to communicate – similar to the ability to act – is seen as a higher-level ability that requires not only the ability to perceive or feel, but also to process and understand. While popular literature often distinguishes according to the medium (e.g., oral, written, etc.), in communication research the distinction according to the number of participants (intrapersonal, interpersonal, transpersonal) is a widely accepted criterion. The influence of feedback on the communication process (unidirectional, bidirectional, omnidirectional) is also often considered a characteristic feature of communication. With regard to the capabilities of AI systems, the focus is currently on enabling human-machine interaction or human-machine communication. Machine-to-machine interactions are currently generally "human-made". In the future, however, it is conceivable that machine-to-machine communication could be enhanced by algorithms to achieve spontaneous, task-oriented, and flexible machine-to-machine interaction.

Table 3: Classification of AI capabilities

Field	Domain	Capability	Examples
PERCEIVE	External	Sight	Optical character recognition (OCR)
			Object recognition
			Gesture recognition
			Infrared vision
		Hearing	Speech recognition
			Audio recognition
			Recognition of radar signals
		Smell	Scent detection
			Acid detection
			Fire detection
			Caprylic acid detection
		Taste	Sugar detection
			Acid detection
			Salt detection
			Bitter detection
			Umami detection
		Touch	Temperature recognition
			Pressure recognition
			Pain recognition
			Electromagnetism recognition
	Internal	Self-perception	Recognition of own movement
			Recognition of body positions
		Balance	Balance recognition

Field	Domain	Capability	Examples
PROCESS	Factual	List	Context-specific terminology
		Summarize	Automated report generation
		Answer	Data association
		Select	Semantic searching
		Check	Parsing (syntactic analysis)
		Generate	Deductive knowledge extraction
	Conceptual	Recognize	Recognize named entities
		Classify	Recognize semantic domains
		Create	Declare
		Differentiate	Clarify terms
		Determine	Semantic interpretation
		Build	Language translation
	Procedural	Remember	Remember processes
		Refine	Model the environment
		Execute	Discourse modelling
		Integrate	Fusion of sensor data
		Evaluate	User modelling
		Design	Model human processing
	Metacognitive	Identify	Select strategies
		Predict	Determine states
		Use	Transfer approaches
		Deconstruct	Coding change strategies
		Reflect	Self-optimization methods
		Create	Narrative generation

Field	Domain	Capability	Examples
ACT	Physical	Movement planning	Movement planning with uncertainty
		Sensors & manipulators	Passive & active sensors
		Kinematics & dynamics	Dynamic movement
		Human-robot interaction	Multimodal teleoperation
	Non-physical	Software agents	Process automation
			Transaction systems
			Chatbots & customer service
COMMUNICATE	Speech processing	Text generation	Paraphrasing tools
		Machine translation	Language translation
		Text analysis	Parsing (syntactic analysis)
		Information & knowledge extraction	Recognize named entities
		Information retrieval	Semantic searching
		Document analysis	Recognize semantic domains
		Speech dialogue systems	Clarification dialogues
			Narrative generation
	Human-machine interaction	Cognitive systems	User modelling
		Interaction paradigms & modalities	Multimodal interaction

Classification of AI applications

The classification of AI applications is often based on the AI methods and capabilities described above. The goal of an AI application is to implement the mathematical methods and abstract capabilities in a concrete way using software. In this way, specialized software markets have emerged to market these typical AI products. These can be purchased or rented by companies and users to increase the productivity of business processes or enable innovations in business models. The typical software markets (see Table 4) are also uniformly designated worldwide and are regularly monitored by independent market analysts (e.g. IDC, Gartner, Forrester, etc.), so that potential users, projects and investors can be well informed about the state of the art.

The software markets can be broadly classified into business intelligence & decision support, AI-based customer interaction, AI-based services and AI development environment & tools.

In business intelligence & decision support, the focus is on creating reports that are timely and on-topic. These aim to ensure a quantitative and qualitative overview of the business and have been commercially available for many years in all areas – e.g. finance, human resources (HR), development, marketing and sales. In this way, decisions are supported and complete planning processes are enabled in complex environments. These capabilities include analytics, as they typically require the analysis of multidimensional data spaces. Key products in this area are software environments for mathematical and AI-based optimization as well as computation of predictions. Another area is the processing of language typically for searching, navigation, and exploration in large bodies of text. Put several of these functions together and entire business processes can be automated, which is often referred to as robotic process automation.

Since 2012, the AI trend has accelerated significantly as available CPUs and GPUs (central and graphics processing units) become more powerful and AI methods based on artificial neural networks can be implemented faster and at lower cost. This allows new possibilities for the human-machine interface, based on AI applications that simulate SMS, chats, speech and physical movements and automate corresponding processes, e.g. simple dialogues in call and service centres.

To simplify the use of AI applications, typical AI applications are offered in public or private cloud environments. This allows users to start customizing the application to their own needs right away, rather than having to spend a lot of time building hardware and software. Typical AI services that are offered out-of-the-box are: Image recognition, video analytics, speech-to-text conversion, text-to-speech conversion, translation, text analytics, intelligent searching, and machine learning. In all of them, the actual use of the artificial neural network is encapsulated and facilitated by a simple graphical user interface or by simple function calls from standard languages (e.g. Java, C, Python, etc.).

Appropriate AI development environments and tools are needed for the development of AI applications. These take into account the typical phases of an AI project: Build, train and run. In all phases open source technologies and software libraries are frequently used, which on the one hand offer AI methods and on the other hand offer professional software development, such as method-based development and development in distributed teams.

By regulating systems based on AI, possible inadequacies of AI applications and competition-distorting constellations can be avoided. In line with the European Commission’s White Paper “On Artificial Intelligence – A European Approach to Excellence and Trust”, the following aspects are important with regard to regulation: Liability, transparency and accountability, as well as training data, retention of data and records, information to be provided, robustness, accuracy, human oversight and specific requirements for certain AI applications, e.g. remote biometric identification applications.

The ethical aspects of the development, benefits and standardization of AI are currently under special discussion. The following characteristics play an important role here, which should be methodically and technically thought through and ensured for each AI application: Autonomy & control, transparency, stability against disturbances, security and all aspects of data protection.

Table 4: Overview of software markets and typical products

Software markets & typical AI applications	
Software market	Typical software products
Business Intelligence & Decision Support Systems	Business Intelligence
	Decision Support
	Work-flow systems
	Planning Analytics
	Constraint Based Optimization
	Prediction Capability
	Text Processing Platforms & Search Engines
	Robotic Process Automation (Rule-Based)
	Cognitive Automation (Training-Based)
	Real-Time Processing
AI-based Customer Interaction	Chatbots
	Voicebots
	Avatars
	Virtual & Augmented Reality
AI-based Services consumed from public or private cloud	Image Recognition
	Video Analytics
	Speech To Text
	Text To Speech
	Translation
	Deep Learning as a Service
	Knowledge Navigation
	Knowledge Exploration
	Intelligent Search
	Natural Language Processing
	Automatical Annotation
AI development environment & tools	Build & Develop AI
	Train & Optimize AI
	Run & Manage AI
	Ethic Support Tools

Classification of AI autonomy
AI applications and the computer systems that implement them can have different degrees of decision autonomy. For example, the German government's Data Ethics Commission distinguishes between three classes of autonomy:

→ Algorithmically based AI applications operate as pure assistance systems without autonomous decision-making authority. However, the (partial) results and (partial) information they calculate form the basis for human decisions.

→ Algorithm-controlled AI applications take partial decisions from humans or shape human decisions through the results they calculate. As a result, the actual decision-making scope of humans and consequently their possibilities for self-determination shrink.

→ Algorithmically-determined AI applications make decisions independently and thus exhibit a high degree of autonomy. Due to the high degree of automation, there is no longer a human decision in individual cases, especially no human review of automated decisions.

4.1.2 Requirements and challenges

4.1.2.1 Ethics

Ethical principles in the context of AI and standardization
An essential task of ethics is to establish and justify generally acceptable standards, oriented to values and principles (e.g., human dignity, justice, freedom), from which instructions for action and behaviour for human (co)life can be derived with a justified (rationally comprehensible) claim to general validity. On this basis, established social moral concepts are also critically scrutinized once again (cf. "Standardization Roadmap Artificial Intelligence Edition 1", Chapter 11.2 "Philosophical Foundations of Ethics", [63]).

Since the beginning of the 20th century at the latest, ethically relevant questions and problems no longer arise solely in the context of interpersonal interaction, but also as a result of the effects of new technologies on human (co)life or in the interaction between humans and technology. Against this background, applied ethics developed as a further sub-discipline from the middle of the 20th century. It deals with corresponding specific aspects that go beyond the classical questions of ethics (e.g. medical, technical and business ethics). Its aim is to bring ethical standards and the norms and principles derived from them to bear in the sense of general rules of the game in corresponding application-specific contexts. AI ethics should also be understood in this context.

Moral principles are not necessarily explicit or even formalized; they can also be implicit conventions of individuals and groups that influence their actions. These general principles have to be transferred by actors to their concrete situation and operationalized in relation to the context of the respective situation. In constitutional states, parts of established rules may be formalized as law, although this is not always congruent with the general moral principles and values of a society, which may also change over time (see Figure 17, right). Taken together, both represent the framework of socially represented goals and expectations, according to which AI must be guided both by morality in its respective context and by applicable law.

An AI system produces effects that operate within and outside a socially negotiated ethical framework (Figure 17, middle). A distinction must be made between the goals explicitly pursued by the AI system and the modalities of goal achievement or implementation aspects (Figure 17, top left). For example, a system can obviously violate ethical values because their fulfilment has not been sufficiently taken into account in the system's objectives. However, even a system that aims to adhere to all relevant moral principles can be criticized from an AI ethics perspective if the modalities for achieving the goals are insufficient; for example, if it is not possible for outsiders to judge whether the system actually adheres to these values, or if the system fails to achieve these goals, for example, due to a lack of robustness.

Accordingly, a distinction can be made between those expectations that are placed on the explicit system goals and those expectations that are placed on the way in which the goals are achieved. Only the consideration of both aspects leads to an overall effect of the system that proves itself in ethical reflection.

Ensuring this is a process that is important throughout the life cycle of an AI system (Figure 17, bottom left) and must be ensured by different actors responsible for the AI system (AI system owners for short, for example clients, developers, test organizations, public bodies or operators). Along the seven phases of this life cycle, as well as in the context of overarching governance, different needs for standardization arise, which are presented and classified accordingly in the chapter on ethics (Chapter 4.1.3).

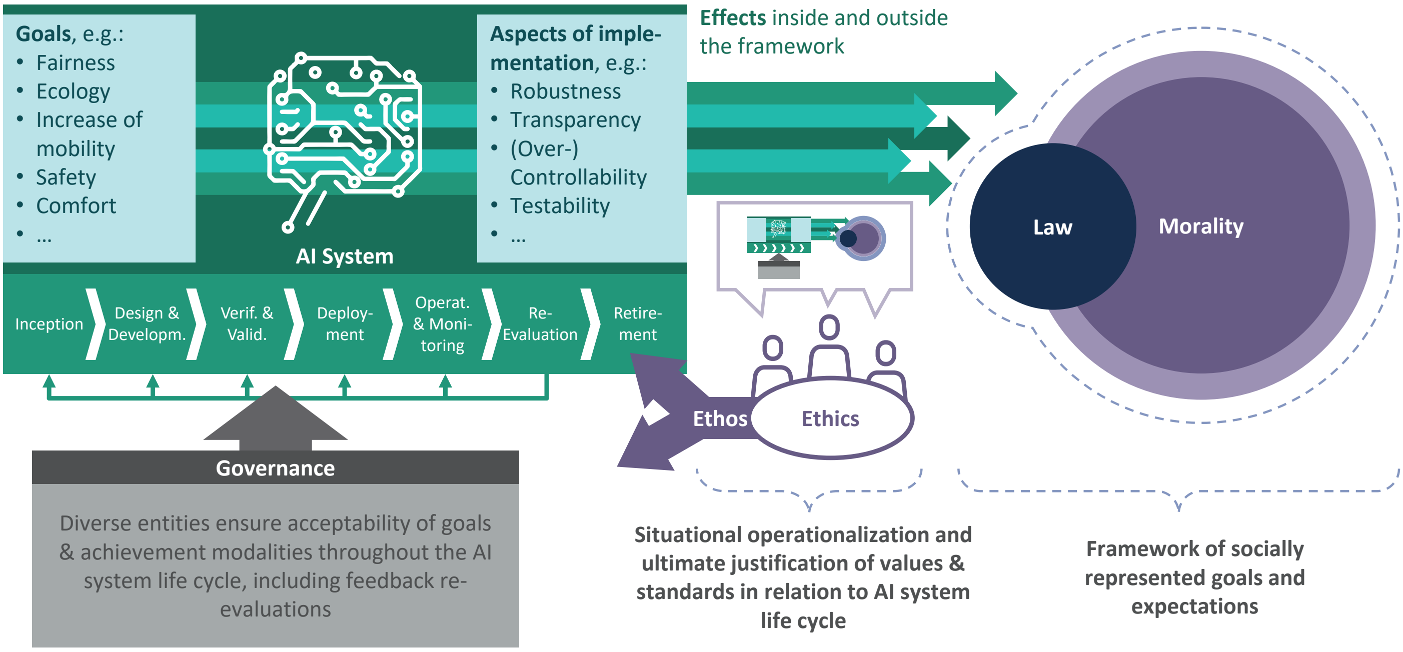

Goals, e.g.:
• Fairness
• Ecology
• Increase of mobility
• Safety
• Comfort
• …
AI System
Aspects of implementation, e.g.:
• Robustness
• Transparency
• (Over-) Controllability
• Testability
• …
Effects inside and outside the framework
Law
Morality
Inception
Design & Developm.
Verif. & Valid.
Deployment
Operat. & Monitoring
Re-Evaluation
Retirement
Ethos
Ethics
Governance
Diverse entities ensure acceptability of goals & achievement modalities throughout the AI system life cycle, including feedback re-evaluations
Situational operationalization and ultimate justification of values & standards in relation to AI system life cycle
Framework of socially represented goals and expectations

Figure 17: Ethics between AI system life cycle (Source: adapted from [16], Working Group Basic topics)

Ethically, the development and operation of AI systems are viewed as to how they operationalize values in concrete terms, i.e.: implement them. Here, it is possible to check how an AI system ensures the principle, for example: “AI systems must respect human self-determination”. It is therefore about the concrete application of AI that is ethically classified and evaluated. The aforementioned persons responsible for the AI system must regularly ensure and comprehensibly explain that the AI system for which they are responsible continuously complies with ethical principles. Design decisions also favour those that promote adherence to ethical principles. Wherever compliance with ethical principles is jeopardized, the responsible actors must explain thoroughly (in the sense of convincingly and rationally comprehensibly) why this is the case and, if necessary, draw the consequences (e.g. conditions for operation, decommissioning or no operating license, etc.).

The actors involved in operationalizing ethical principles in the environment of a specific AI system can be diverse. However, it must always be assumed that those designated as being responsible for the AI system are significantly involved in the ethical considerations related to their AI system. This is because they have a duty to bring the AI system for which they are responsible into line with the legal and identified ethical principles in its concrete context of action.

Those responsible for AI systems can promote and justify the value-based development and operation of the systems through ethical reflection (see also [64]). This apparent value base is a key factor in AI systems gaining social acceptance. Sensitivity to and criticality of the technical implementation of a value base, for example, are two aspects that shape the sense of duty of those responsible for AI systems. This sense of duty is comparable to a professional ethos of those responsible for AI systems (analogous to the Hippocratic oath in medicine). It is clear that the aforementioned ethical reflection in the development and operation of AI systems strengthens and updates this ethos.

Ethical reflection in the development and operation of AI systems can involve the following basic steps, which will be illustrated here using the field of medicine as an example. The example represents an AI-based diagnostic system for skin cancer detection as a smartphone application, a teledermatology app (referred to below as the Derma app). Users photograph the relevant skin area. The Derma app analyzes the photograph and makes a recommendation. In case of a possible suspicion of skin cancer, it advises the user to consult a specialist doctor.

Aligning AI development and operations with ethical principles and values requires the following steps:

Those responsible for the AI system	Example Derma app: AI-assisted diagnostic system for skin cancer detection as a smartphone application
→ develop an AI design based on values: They define the understanding of values (for example, based on their code of conduct) and prioritize values for their use case, i.e., including the respective context. They make this process transparent and comprehensible.	For the Derma app, those responsible for the AI system take into account various values and ethical principles, for example: → **Equal treatment** and **explainability**, for example by or for people with different educational backgrounds: by means of clear model representations or explanations in easy-to-understand language, it must be reasonably comprehensible how the app works and on what basis it makes recommendations. Users must be able to classify the recommendation as a guide and not as a substitute for treatment by medical professionals. It must be clear that a holistic diagnosis also includes other examinations (e.g., a palpation result) in addition to the visual observation, as well as a progress and comparative observation. The app cannot do this. Equal treatment also implies that the use of the app (e.g., in terms of taking pictures or entering further data) can be reliably implemented for different user groups without causing significant restrictions in the reliability of the results (see below re diversity). → **Reliability**, here in terms of the clarity of the basis for the recommendation: In order for patients and doctors to be able to assess the basis on which the Derma app makes recommendations and the likelihood that these recommendations are correct, they need concrete insight. For example, the parts of the skin photograph that are decisive for the diagnosis can be marked, so that users can directly see which features the app uses to make its recommendation. A continual improvement in the rate of accurate recommendations is also required as part of quality assurance. This can also be achieved by including various training data (see below). To reduce misdiagnoses, those responsible for the AI system can provide for balancing of artefact sources (so that, for example, skin photography exposure errors are not misinterpreted as pathogenic abnormalities). → **Diversity**, for example, through appropriately diverse training data: The app must be able to be used equally by all people (regardless of age, gender or skin colour). To avoid overfitting to specific patterns, those responsible for the AI system can use data from multiple sources, such as different laboratories. → **Self-determination**, e.g. with regard to a suitable user interface with corresponding intervention options: Dermatologists and patients need to be able to provide feedback to those responsible for the AI system. It must be possible for those responsible for the AI system to react to this in order to eliminate sources of error. They must be able to delete incorrect or erroneous data or reset the AI system. To incorporate and prioritize values to the extent necessary, those responsible for the AI system engage in dialogue with stakeholders. Together with representatives, for example from the groups healthcare professionals and patients, they examine the listed values and their prioritization, taking into account diversity aspects (such as age, educational background, gender, etc.). They document this result and make it publicly available in its most important aspects and in a clear form accompanying information from the Derma app.

Those responsible for the AI system	Example Derma app: AI-assisted diagnostic system for skin cancer detection as a smartphone application
→ formulate requirements for the AI system based on their findings on the relevant values: Based on the respective target value and the application context of the system, they determine key requirements on how its functions are to be implemented in compliance with the created value listing and prioritization. They take a systematic approach to align the evaluation of individual sub-requirements for their AI system and achieve „ethics by design."	Two values that are exemplified as mutually reinforcing for the above use case are **equal treatment** and **explainability**. It can be argued that both pay into the target value of **self-determination**, because a fundamental understanding of the functions and processes of the Derma app is required for well-founded criticism and feedback capability on the part of affected groups. In order to implement the ethical principle "AI systems must respect human self-determination" as an overriding goal, those responsible for the AI system must ensure that users retain sovereignty over their decisions at all times. But also the understanding of a recommendation of the system must be given in a sufficient way with different user groups. If such understanding is limited for certain user groups, e.g., people with insufficient technology or media literacy, those responsible for the AI system must specify measures that can ensure appropriate assessment. This could be achieved, for example, through the mandatory involvement of additional persons such as medical professionals: A dialogue area can be set up within the Derma app where patients can contact doctors with their queries. Similarly, there may be an area to provide further contact with medical practices and medical outreach clinics to achieve immediate networking and assistance.
→ must describe and resolve conflicting goals in terms of their values.	There may be different trade-offs with respect to the above-mentioned value of **reliability**. The threshold value at which the Derma app recommends a visit to a doctor should be taken into account at this point: → If the threshold value is very high, the Derma app would react very sensitively to skin changes. A doctor's visit would then tend to be recommended frequently (risk of false positives). → If the threshold value is very low, the Derma app would react comparatively insensitively to skin changes. A doctor's visit would then tend to be recommended less often (risk of false negatives). In terms of reliability, it is now necessary to consider: Are those responsible for the AI system to accept more false positives to reduce the risk of missed alarms and to enable early detection that is reliable in this sense? Or are they to accept more false negatives in order not to overburden the health care system with unnecessary examinations and treatments as well as burdens for users? It is important to balance these two aspects. In addition to the balancing of objectives within a value, there is also the question of balancing two values in opposition. At this point, **equal treatment** (here: regarding access to the system) should be balanced with **reliability** (here: through quality assurance). If the same access is granted for all user groups, the reliability of the app may be reduced for less technically savvy people if the required recordings cannot be created properly or the results cannot be interpreted correctly. This concerns, for example, different age groups and different technical, and also linguistic understanding. In this respect, the value of equal treatment competes with the value of reliability, and a balance must be achieved between these aspects as well, for example, by designing the user interface in such a way that it is accessible to different user groups in an appropriate manner or is safeguarded by appropriate measures (e.g., involvement of specialist personnel).

Those responsible for the AI system	Example Derma app: AI-assisted diagnostic system for skin cancer detection as a smartphone application
→ demonstrate whether the AI system ultimately functions according to the identified requirements and ensure ongoing quality assurance.	As the above shows, those responsible for the AI system for the Derma app must determine how individual values are implemented in a traceable manner. They determine through which measures... → users have equal access to the app, → recommendations are classified appropriately and in a user-friendly manner, → recommendations are ensured with reasonable reliability, → users can use the app in a self-determined manner, and how this can be verified. It should be noted that there are different levels of verification or validation. Ultimately, those responsible for an AI system must examine not only the individual requirements, but the system as a whole. Only in this way can they assess the interactions of individual components or decisions and the potentially resulting conflicts of objectives, especially in complex systems. This concerns an overall evaluation that integrates clinical effectiveness criteria and ethical aspects. In essence, it must be validated whether the values specified at the beginning of the development could be implemented in the present use case in a sufficient manner. This includes representative coverage of the user groups and application contexts present in the use case. Since it is often not possible to cover and/or foresee all situations during the development process, it is also necessary to systematically collect data from the operation of the system in the sense of a quality backward chain and to have it re-evaluated at regular intervals by a suitable body. In the present example, the quality backward chain would include a systematic review ... → of the extent to which the individual user groups have received the right decisions for their personal case, or → whether the human oversight mechanisms (involving other people) were effective enough to handle this individual case appropriately. The re-evaluation is a review of whether the ethical dimension of the envisaged objectives could be implemented appropriately across the entire spectrum of use cases, or of where there is a need for action with regard to improving the system or the associated business processes. Important aspects include: → the avoidance of potentially systematic unequal treatment of certain user groups and promotion of equal treatment; or → the reliability and explainability, i.e. the accuracy with which the app makes recommendations and presents them in a way that users can understand. This is the basis for trust in their functionality on the part of all target groups.

Standardization can support the complex process of implementing values in the development and operation of AI. Standardization ...

→ provides impetus for goals that are suitable for justifying the ethical defensibility of an AI system. In doing so, it takes up the central ethically relevant questions and problems of this special field, which are identified socio-politically,

→ provides the basis for arguments that can be used by people acting ethically as part of their discourse,

→ creates an intersubjective understanding of language by shaping and defining terms (the use of a common language is what makes communication and the exchange of arguments possible in the first place),

→ develops schemas to classify AI systems in a uniform way,

→ develops procedures that standardize ethical processes and make value-based system requirements measurable.

Standardization supports the implementation of values – sometimes making thinking, communicating and arguing with regard to ethically relevant issues more efficient in the context of AI. A key goal here is to lay the foundation for developing and operating AI in a systematic and contextually trustworthy manner – that is, in terms of the value trustworthiness. The following section deals with this aspect and its prerequisites.

Value systems for trustworthy AI

The term "trustworthiness" can basically refer to both organizations and technical systems. In contrast, it is to be specified that: Ethics [65] only refers to "rational beings" who are actors (e.g. those responsible for AI systems) in organizations, but not in technical or algorithmic systems. More concrete explanations of trustworthiness in relation to organizations or technical systems can be found in Chapter 4.1.2.2.

Values and requirements for trustworthy AI in general

The "High Level Expert Group on Artificial Intelligence of the European Commission" (HLEG-AI) [8] as well as the "Enquete Commission AI" [66] have described a number of requirements for AI systems with regard to their trustworthiness. These values or requirements for trustworthy AI systems, referred to as guidelines, include the following points (see Chapter 1.4).

1. The prioritization of human oversight of AI systems and respecting and ensuring fundamental rights: The Group requires that information, oversight and control mechanisms should be available in connection with AI systems in order to avoid negative effects, e.g. on fundamental rights, but also the misuse of AI systems.
2. Technical robustness and security, e.g. resistance to attacks and security breaches, fallback plans and general security, precision, reliability and reproducibility.
3. Privacy and data quality management, such as respect for privacy, data quality and integrity, and data access. Issues that relate to standardization activities include data protection management in the context of AI, but also how to ensure data quality overall.
4. Transparency, traceability and explainability. In practice, these terms are often used synonymously. However, they relate to various aspects of disclosure, as defined below.
   a) Transparency refers to the question of „what". It aims to make the use of AI components in a system recognizable and to describe the system's relevant properties. This knowledge is necessary to enable an informed decision about the use of the AI system.
   b) Traceability in this context refers to the possibility of being able to independently verify the properties made transparent.
   c) Explainability refers to the question of „why." Through explainability, the behaviour of AI components and their interaction in a concrete situation can be understood. This knowledge makes it possible to trace decisions of the AI system back to their influencing factors and thus to understand the cause of individual decisions. Datasets and processes that led to the AI system decision should be documented.
5. Fairness, non-discrimination, and diversity, e.g., avoiding unfair bias, accessibility and universal design and stakeholder participation, promoting diversity.
6. Social and environmental well-being, e.g. sustainability and environmental protection, social impact, society and democracy.
7. Accountability, e.g., verifiability, minimizing and reporting negative impacts, compromises, and remedies.

This can also be compared to the Landscape of AI Ethics guidelines, which identified five values or ethical principles of fundamental importance for AI systems: Transparency, justice, fairness, non-maleficence, responsibility and privacy [67]. Another approach to value-based development and the use of AI systems can also be found, for example, in the White Paper Ethics briefing [68].

The value of fairness in particular

Fairness has a special position as a requirement (see item 5 above) for trustworthy AI systems for several reasons. On the one hand, society rightly demands fairness in general and in principle, especially in exponential technologies such as the application of artificial intelligence; on the other hand, fairness as an operationalization of non-discrimination (in the sense of unjust disadvantage, bias, or unequal treatment) has already become established in computer science and its practical application in the last ten years.

If the general definition of fairness according to the German Duden dictionary of “decent behaviour; fair, honest attitude towards others” or “corresponding to the [game] rules, decent and comradely behaviour in the game, competition or the like” is widely accepted, a common more specific definition, for example, from the two perspectives of the disciplines of philosophy and technology would already be much more difficult. Even limiting the focus to just one discipline, such as computer science, is still a challenge when defining fairness specifically.

However, as fairness is increasingly required in the use of algorithmic and sociotechnical systems more broadly and machine learning systems more narrowly, action is required. The meaning of the term in this context is also highly controversial. Broadly speaking, two main streams can be distinguished: Fairness as an ethical principle (based on values such as justice) and fairness as an operationalization of non-discrimination. It is often not clear what the call for fairness is based on in a specific case. In terms of operationalizing non-discrimination, however, there are not only concrete implementation strategies, but also already concrete proposals for measurement and assessment that are used in practice

Over the past decade, a limited understanding of fairness in computer science has developed in parallel with an ethical understanding from applied philosophy. In computer science, there is an effort to measure “only” the extent of discrimination by an algorithmic system through fairness measures inversely (i.e., the “non-discrimination”). This does not cover all aspects of fairness.

The variety of approaches to measuring fairness in terms of non-discrimination represent different perspectives and strategies and can be broadly divided into individual and group fairness measures. In any case, a universally applicable fairness measure presupposes a common understanding of discrimination. However, this is not available or given by various moral concepts, systems of norms, principles, values or dispositions, all of which claim to be the basis of correct action (see Glossary “Ethics”). For an ethical reflection it is furthermore essential to examine discrimination occurring in AI-supported applications (cf. [69], 3) as a possible extension of social inequalities embodied by humans as social actors (cf. [69], 6). Therefore, the social background, structured by possible hierarchical power asymmetries, from which algorithmic systems can emerge, should also be taken into account (cf. [70], 2). Thus, there cannot be “the one” fairness measure, but rather a deliberate selection of fairness measures should be made to measurably and demonstrably promote the intended fairness goals. In this respect, it appears essential – in order to be able to do justice to the value of fairness in its respective context-related implementation – to enter into dialogue with relevant stakeholder groups (as also exemplified in the example under item 1 of this section) in order to be able to determine and take into account the fairness opportunities and challenges of an AI system in direct relation to the respective stakeholders. In addition to value-based engineering in the context of the above-mentioned IEEE 7000:2021 [64], approaches such as participatory design [71] or value-sensitive design [72] also incorporate this aspect. There is disagreement about group fairness measures in that they must be about the (conditional) equal treatment of groups. Individual fairness, on the other hand, is the view that similar individuals should be treated similarly, based on some (arbitrary) function that determines similarity. It should be noted here that unequal treatment may also be justified (e.g., in the allocation of a job that requires a high level of physical strength, or the prioritization of vulnerable groups in the allocation of vaccines).

Different measures of fairness represent different notions of fairness, and many of them cannot be optimized at the same time because they conflict with each other to some degree. If optimization is targeted to a specific fairness measure, the results of other fairness measures are sometimes inevitably reduced. This may even increase discrimination according to the understanding of reduced measures (see Chapter 4.8.2.3).

Since it is virtually impossible to map morally imperative actions in fixed algorithms or rigid sets of rules, a trustworthy organization consisting of “rational beings”, i.e. employees, (loosely based on Kant) is characterized by its ability to behave in an ethically reflective manner, especially in conflict situations, even if existing laws or company regulations might conflict with this (see Ethical Guidelines of the Gesellschaft für Informatik e. V. (Society for Informatics) [73]). Modern governance and management systems (see Chapter 4.1.2.2) include clear and effective compliance reporting channels for precisely such conflict situations to protect employees.

Case example governance
The possibilities for implementing values such as trustworthiness will be described below using an example based on the specific case of a large software company that has been in practical use for several years.

In the example, the operationalization in the form of “principles” for an ethical approach to AI goes back to an initiative of the employees. They enlist the support of top management and conduct international workshops with global participation from all business units affected by AI or ML. The principles developed herein include three perspectives or roles: Employees/employers, solution providers and members of society. The principles describe their interaction in accordance with the principle of sustainability, in the sense of the conscious use of tangible and intangible resources in such a way that their creation, use and further development today do not compromise the needs of future generations.

These abstract principles are then concretized into “guiding principles” and further detailed into instructions for action and rules, for example as follows:

Principle	We develop for people. (This goes back to Kant’s self-purpose formula and implies, among other things: Technology is always there for people – never the other way around.)
Guiding principle	Clarification for employees on how ethical principles are to be incorporated into everyday working life.
Concrete instructions for action	→ No grey, dark, or black patterns (for example, purposefully misleading user interaction, e.g., with cookie selection options by highlighting or darkening the buttons). → Supply chain check for third-party service providers → No de-anonymization → ...

Concrete tools such as a criticality pyramid or risk matrix (cf. Roadmap AI 1st edition [63]) for classifying the company’s internal algorithmic systems support a comprehensible and low-threshold implementation.

In addition, an AI ethics governance structure consisting of external and internal experts is to be established, for example in the form of an “AI Ethics Steering Committee”, “AI Ethics Office” or “External Advisory Panel on AI”. This structure is responsible for the permanent design and further development of the principles, guiding principles and recommendations for action, inherently maps the corporate values and keeps them up to date.

This example shows practical steps that can be taken within the company to operationalize ethics. However, there are already efforts by organizations and academia to provide cross-enterprise process structures and concepts in this area (e.g., IEEE 7000:2021 [64] and KIDD process [74]). Standardization can support here to provide a reference and ensure the comparability of measures.

4.1.2.2 Implementation in AI development and operations: A look at products and services as well as organizational structures

As shown in Chapter 4.1.2.1, the term “trustworthiness” can refer to both organizations and technical systems. A technical system (i.e., a product or a service provided electronically) can be trusted with respect to certain properties such as safety/security or reliability if there is evidence (e.g., in the form of a test report or certificate) that the system meets such properties. The trustworthiness of an organization is broader: It refers to the fact that an organization is trusted to implement appropriate actions and maintain management structures – called a management system – to meet the expectations of its stakeholders and other interested parties. In addition to an appropriate audit report, an organization’s reputation or its acceptance in the marketplace can also contribute to its trustworthiness.

Trust in products and services
The Common Criteria (CC) describe a methodology for testing products and services with a focus on their safety, which can be used as a conceptual framework for corresponding tests of AI systems. The CC are also available as International Standard ISO/IEC 15408:2020 [445]. Supporting this, an agreed methodology for evaluation based on the CC is described in the International Standard DIN EN ISO/IEC 18045:2021 [75]. These documents form the technical basis of the Common Criteria Recognition Arrangement (CCRA) [76], which has been signed by a large number of countries, including Germany.

Further information on the CC can be found, among other places, on the BSI website [77].

Requirements for an examination according to the CC are summarized in Evaluation Assurance Levels (EAL):

EAL1	functionally tested
EAL2	structurally tested
EAL3	methodically tested and checked
EAL4	methodically designed, tested and reviewed
EAL5	semi-formally designed and tested
EAL6	semi-formally verified design and tested
EAL7	formally verified design and tested

Trust in organizations

For further investigation of the AI HLEG requirements on the trustworthiness of AIs, a conceptual digression will be undertaken to distinguish between the terms "governance" and "management", as is currently done in ISO/IEC 38500:2015 [78]. It should be noted that the term "management system", refers to all three levels discussed in the following, namely the governance body, the management and concrete technical and organizational measures, as shown in Figure 18.

Governance

Governance refers to the general tasks and objectives of an organization, its self-image and the resulting values and culture of the organization that determine its actions. This includes, in particular, the ethical value system to which the organization subscribes (see 4.1.2.1 Ethics). A central concept is that of risk-taking. According to the conceptual framework of ISO/IEC 38500:2015 [78], the governance body of an organization is responsible for implementing its accountability and due diligence obligations. Liability issues in particular also take on special relevance in connection with AI, since the possible degree of automation of AI makes the question of who is liable in the event of errors and damage important. Governance should take this into account, as the legal framework in this field is dynamically evolving. The governance body provides specifications and establishes guidelines that must be implemented within the organization.

Furthermore, the governance body is responsible for establishing management structures (processes, roles, responsibilities) and providing adequate resources.

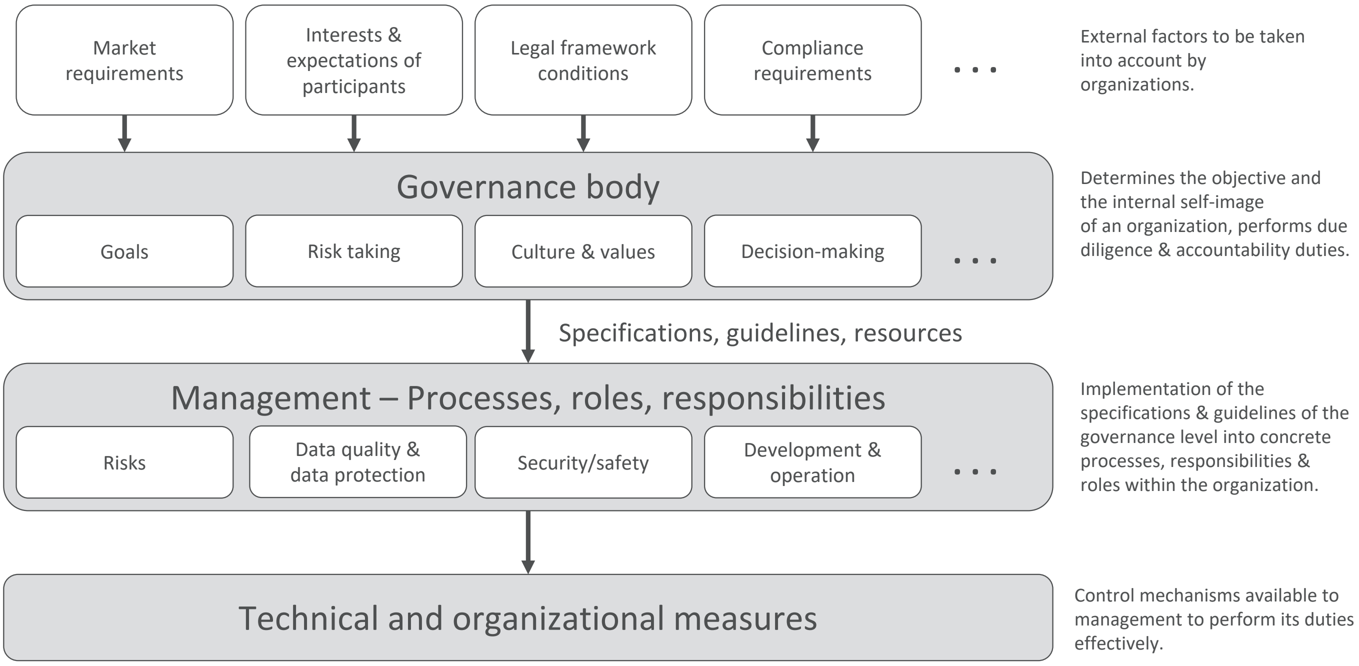

Market requirements
Interests & expectations of participants
Legal framework conditions
Compliance requirements
. . .
External factors to be taken into account by organizations.
Governance body
Goals
Risk taking
Culture & values
Decision-making
. . .
Determines the objective and the internal self-image of an organization, performs due diligence & accountability duties.
Specifications, guidelines, resources
Management – Processes, roles, responsibilities
Risks
Data quality & data protection
Security/safety
Development & operation
. . .
Implementation of the specifications & guidelines of the governance level into concrete processes, responsibilities & roles within the organization.
Technical and organizational measures
Control mechanisms available to management to perform its duties effectively.

Figure 18: Management system: Governance, management and technical-organizational measures (Source: Peter Deussen)

Management

The management of an organization translates the provisions and guidelines of the governance body into concrete processes, roles and responsibilities. Examples of management tasks include:

→ the identification and analysis of potential risks and the establishment of options for action based on the willingness of the organization to take risks,
→ the establishment of a data protection management system and processes to ensure sufficient data quality,
→ the introduction of security management for AI-based IT systems,
→ effective management of the development and operation of AI systems.

Technical-organizational measures

This term covers all technical and organizational tools available to management to fulfil its tasks effectively and verifiably. Technical-organizational measures range from the availability of encryption functions to increase data security to the application of statistical methods to identify unfair bias or contamination in datasets and the availability of test and validation tools.

Requirements on the management system

The concept of the management system standard (MSS) plays a central role in the context of international standardization. An MSS defines requirements for organizations for implementing effective and responsible management. In some cases, requirements are also placed on the governance body of an organization, and many MSS still contain specific controls in the sense of technical and organizational measures. The term "management system" thus refers to the overall picture presented in Figure 18. Minimum requirements for the management system are described in the High Level Structure (HLS) [263].

1. **Context of the organization,** This includes, among other things, the legal framework, social expectations, needs and expectations of interested parties, goals and values of the organization and the actual scope of the management system.
2. **Leadership,** The governance body must define binding readiness of the organization and lay it down in the form of provisions. These provisions include those dealing with the ethical value system. The body must also define processes, roles and responsibilities for effective management.
3. **Planning** includes activities to deal with risks and opportunities.
4. **Support** includes the provision of resources, the determination of necessary competencies, ensuring necessary mindfulness, communication and documentation.
5. **Operation** is the operational implementation of management requirements.
6. **Performance evaluation** comprises monitoring, analysis and evaluation, internal auditing and management review.
7. **Improvement** deals with the identification of nonconformity with regard to MSS requirements, corrective measures and the continual improvement of the management system.

Organizations can demonstrate compliance with MSS (e.g., through self-assessment or certification by an independent third party), thereby increasing the organization's trustworthiness as regards the specific aspects of MSS. When considering the use of a class of technologies such as AI, an organization's management system must therefore refer to the specific characteristics and range of impact of AI. This can be done by extending existing MSS to include AI-specific requirements. However, since the various MSS are published and maintained by different bodies within ISO and IEC that have neither a common conceptual framework nor a synchronized way of working, and furthermore since it is not clear whether existing MSS are even sufficient to address all aspects of AI, it would be more promising to design a new MSS that focuses on AI-specific requirements.

Supporting standards

MSS only include requirements for a management system, but do not describe its implementation. This allows organizations to define their own management structures in the way that suits them, as long as evidence can be provided that the MSS requirements are met. Such structures, but also underlying technical and organizational measures are usually described in supplementary specifications, which do not contain requirements but only guidelines.

4.1.2.3 Development of AI systems

Software gives machines an ever-increasing range of functions. Hardware and software form a symbiosis in this. For software with a predetermined functional sequence, there are generally accepted development and quality assurance procedures, for example code reading, module and application tests at various integration levels, verification and validation. These methods and procedures also apply to software with rule-based AI systems. In addition to the quality of the software code and the compilers used, the software architecture, the quality of the data used and the learning phase are of particular importance when developing AI systems.

Learning AI systems receive essential functionalities through the learning phase. This learning phase can be static or dynamic, supervised or unsupervised. As with humans, the testing of what has been learned is a great and new challenge for software development. This is especially critical because AI systems show their strength especially where decisions or decision recommendations based on a large amount of data have to be made very promptly.

If AI systems are used for automated or autonomous decision-making in safety-critical areas, related procedures for verification and conformity assessment by third parties are required. This applies in particular to evidence within the framework of the proof of functional safety in product liability.

An appropriate approach to the development of AI systems is one considering the entire life cycle of an AI system in its application environment, as well as ensuring data quality in the learning and application phases.

The life cycle of an AI system

The International Standard ISO/IEC 22989:2022 [16] describes a generic life cycle model for AI systems that includes the following phases (cf. Figure 19):

- → Inception: Initial phase of the development process of an AI system in which the essential requirements and design parameters for the project are defined.
- → Design and development: Construction phase of the AI system, in which a functional version is made available for the following phase of verification and validation.
- → Verification and validation: Testing of the AI system with respect to requirements and the fulfilment of project objectives.
- → Deployment: The AI system is installed in its operational environment. This phase includes further testing to ensure that the system operates satisfactorily in this environment.
- → Operation and monitoring: The system is commissioned and monitored during operation.

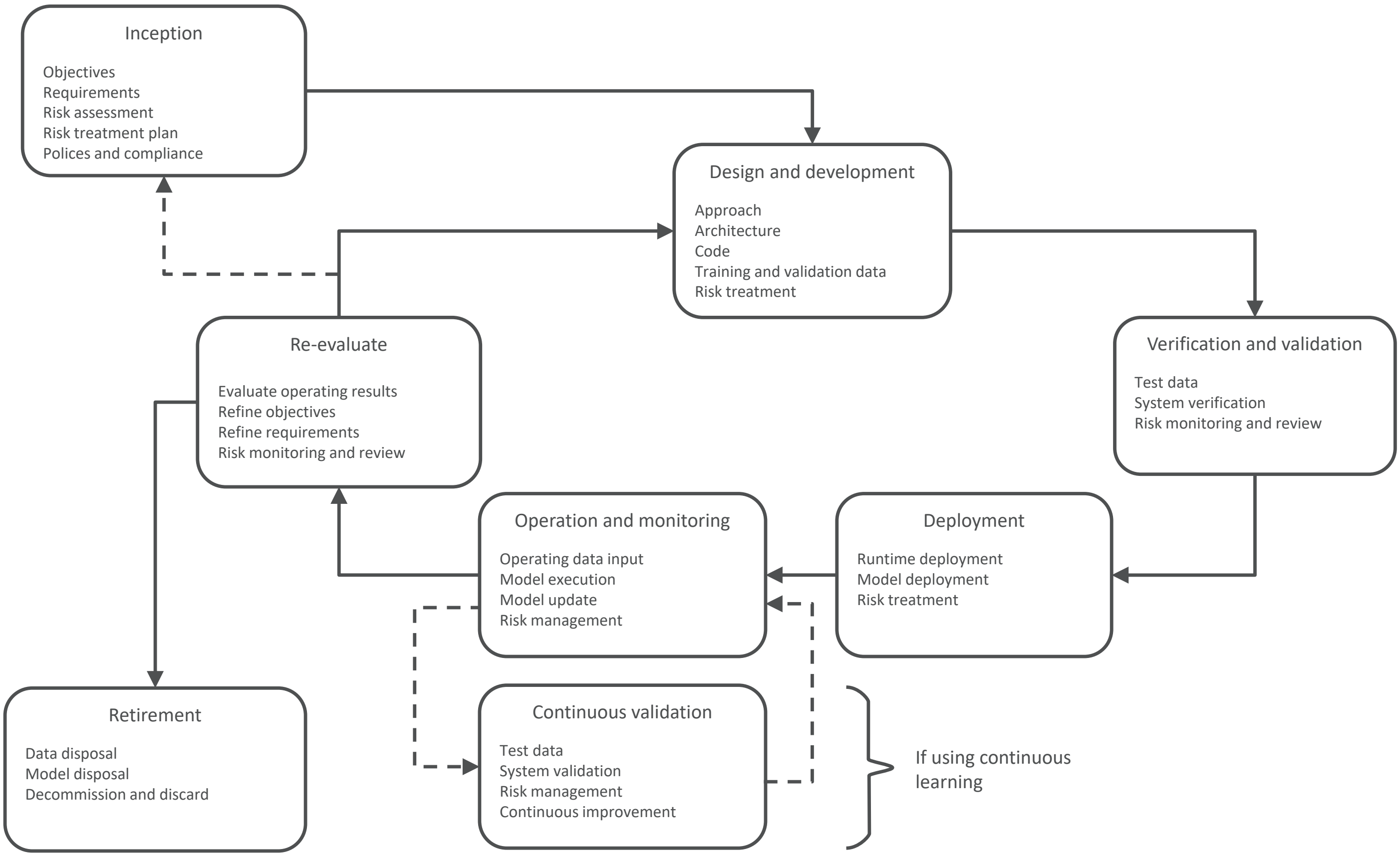


Figure 19: Life cycle for AI systems (Source: adapted from [16])

→ Continual validation: AI systems that continuously adapt to changing circumstances in their operating environment – e.g., through continual learning – must be tested either continuously or at set intervals according to their continued function.

→ Re-evaluation: In longer phases, a re-evaluation of the AI system is performed with respect to changed goals or requirements.

→ Retirement: The AI system is decommissioned.

→ These phases are not interdependent in the sense of a linear sequence, but must be passed through in an interlocked manner. Re-entry into an already completed phase is possible.

Data quality

Analogous to using a system life cycle to organize quality and risk management activities along the development and operational phases of an AI system, a data life cycle model can be used to describe data quality management. The ISO/IEC 5259 series of standards [39], currently under development in ISO/IEC JTC 1/SC 42, addresses data quality management.

Figure 20 relates the data life cycle to data quality management. The phases of the data quality management life cycle include:

1. **Data motivation and conceptualization** Based on the intended use of data, concepts for data management are derived that take into account the relevance of the data, compliance requirements and, where applicable, ethical requirements.
2. **Data specification** comprises the description of required data, usable data formats, identification of incorrect or contradictory requirements of the specification.
3. **Data planning** comprises the planning of the implementation of the data specification including the planning of specific tasks for data acquisition and processing and the provision of the necessary resources for this.
4. **Data acquisition** comprises the collection of data, if necessary in the case of synthetic data, their generation, and the combination with existing data.
5. **Data preprocessing** comprises activities such as cleaning and filtering of raw data or reduction of data volume.
6. **Data augmentation** includes adding metadata to data, categorizing data (labelling), etc.
7. **Data provisioning** involves the use of data for its intended purpose, e.g., for learning a neural network.
8. **Data decommissioning** comprises the deletion of data or the transfer of project-related data to a general database or a new project.

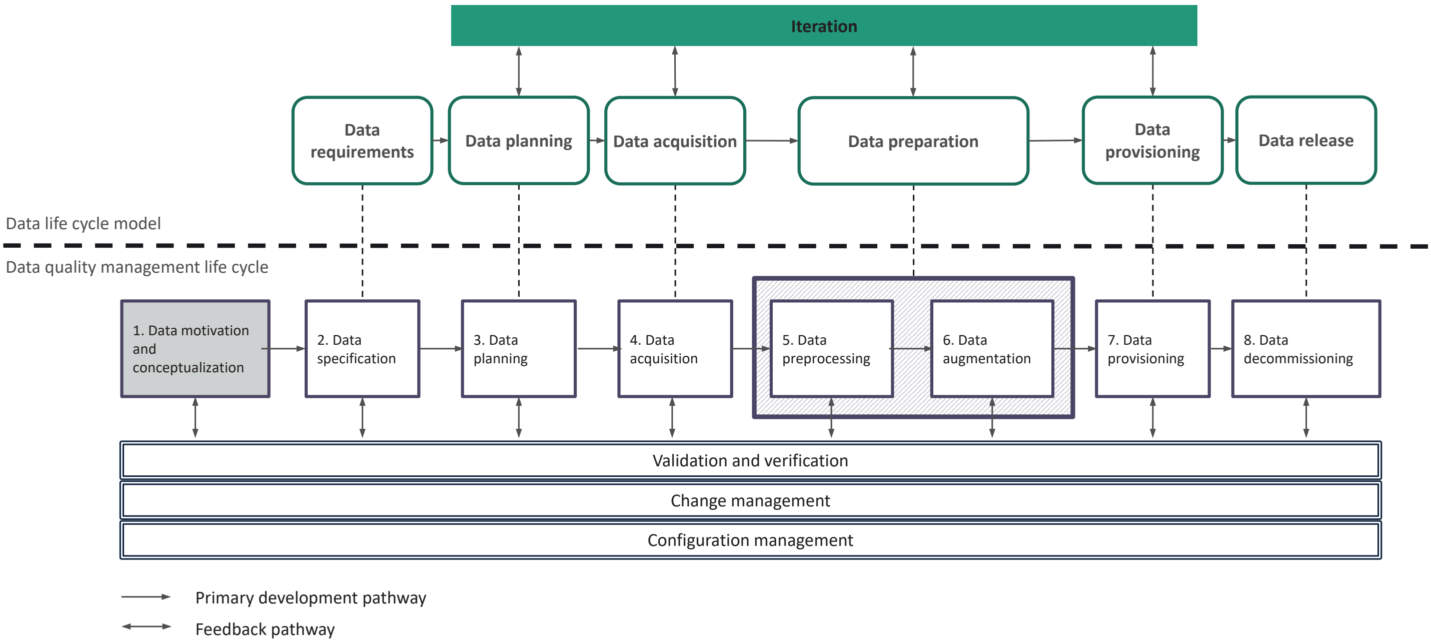


Figure 20: Data life cycle and data quality management life cycle (Source: along the lines of [39])

Quality criteria for data
Quality criteria for data are discussed in International Standard ISO/IEC 5259-2 [41]. This International Standard describes a total of 19 quality characteristics for data:

1. **Portability**: Transferability of data from one system to another.
2. **Comprehensibility**: Degree of comprehensibility of data for the user.
3. **Auditability**: Availability of data for internal or external audits.
4. **Identifiability**: Degree of identifiability of persons with whom data can be associated.
5. **Currentness**: Degree of timeliness of data.
6. **Credibility**: The degree of trust that a user can place in the truthfulness of data.
7. **Completeness**: Degree of coverage of expected information by a dataset.
8. **Scalability**: Degree to which data quality is maintained when the amount of data or data input speed is increased.
9. **Generalizability**: Degree to which data can be used in a context for which it was not originally collected.
10. **Effectiveness**: Degree to which data meet certain requirements.
11. **Accuracy**: Degree to which data correctly reflect a particular fact.
12. **Precision**: Degree of precision to which data make a particular fact distinguishable from other facts.
13. **Consistency**: Degree of consistency of data.
14. **Relevance**: Degree of adequacy of data for a particular purpose.
15. **Timeliness**: Degree of delay in data availability with respect to the time of its collection.
16. **Representativeness**: Degree to which data describe all relevant aspects of a given issue.
17. **Balance**: Degree to which all relevant aspects of a given issue are described by sufficient amounts of data.
18. **Similarity**: Degree to which relevant facts are described by similarly structured data.
19. **Diversity**: Degree of multiplicity of data.

Recommendations for action
With the initiation of the ISO/IEC 5259 series of standards [39], the topics of data quality and data management are being addressed, at least generally, in international standardization. Nevertheless, it is to be expected that for specific sectors and applications more stringent quality criteria, possibly deviating from those mentioned above, will become relevant. Quality management processes must also be implemented and, if necessary, enriched on a sector-specific basis. Thus, it is recommended to examine in vertical standardization on data quality management to what extent the ISO/IEC 5259 series [39] can be used as a general reference and to what extent sector-specific adaptations are necessary.

4.1.2.4 Quantum AI

Modern machine learning (ML) methods are often extremely resource-intensive, especially during their development, but are also still unable to solve certain challenging problems, or at least not efficiently. Quantum computers have the potential here to overcome these limitations.

In this context, the field of Quantum Machine Learning (QML) has established itself as a discipline in its own right, combining approaches of machine learning and quantum information processing (see e.g. [79], [80]). From the perspective of ML developers and researchers, the use of quantum algorithms as part of the classical ML life cycle, especially during the training phase, is a particularly relevant starting point. The basic idea that may be able to resolve the limitations of classical ML methods mentioned at the beginning is to offload certain subprocesses and computations to quantum hardware.

QML is currently a very dynamic field of research where many questions are still open, especially with regard to the practicality of the discussed methods. Nevertheless, significant progress can be expected here in the next few years, especially due to the rapid developments in the field of quantum computing. A scenario in which ML practice is permanently changed by the use of quantum computers should therefore already be discussed now. The opportunities, but also the risks, that QML may bring with it must be considered in detail.

Especially in the area of IT security, many issues arise that need to be addressed with foresight [81]. Two aspects are essential in this discussion: First, it is already known that attacks on classical ML systems [83] (keyword: Adversarial Machine Learning) are in principle also transferable to QML systems. However, the extent to which QML has a higher vulnerability here or can offer improved resilience is still unclear. Second, applications in IT security that currently use conventional ML methods may benefit from any efficiency gains from the use of quantum computers. This basically concerns the perspectives of both the attacker and the defender.

Accordingly, considerable research efforts are still required in order to establish QML as a secure technology from the outset on the one hand, and to assess and address the impact on IT security itself on the other.

4.1.2.5 Language technologies

Language technology is an interdisciplinary field that draws primarily on computer science, artificial intelligence research, and computational linguistics to develop applications, methods, and solutions for the analysis or generation of written or spoken language, with multimodality currently also playing an important role, such as the simultaneous processing of speech and visual data.

A central feature of language technology concerns the fact that it encompasses a wide range of applications: from the classic use case of machine translation (written or spoken language), the spectrum extends to the synthesis of spoken language (e.g., natural-sounding announcements on train platforms) or the generation of written language (e.g. automatic generation of product descriptions), spoken language recognition (e.g., recognition and transcription of text messages on the phone), and written language analysis (e.g., text classification, information extraction, knowledge graph generation, text summarization, entity recognition, syntactic or semantic parsing, etc.). Language is also being increasingly used as a channel for human-machine interaction, e.g., for question-answering systems, information retrieval and search engines, for chatbots, and for smart personal assistants, as they have been available for several years in all modern operating systems, telephones, and various household appliances and also cars.
The historical core of language technology is computational linguistics as well as language-processing, knowledge-based AI, which since the 1970s used in particular manually developed rule systems and symbol-processing methods (semantic networks, taxonomies, ontologies, knowledge graphs).

After a renewed scientific turn to statistical AI methods in the late 1990s, correlative (non-deterministic) neural networks have dominated for the past decade or so, with this development also spurred by the favourable availability of powerful GPUs. Machine learning methods which belong to the field of deep learning dominate science and technology in numerous subfields of language technology, using large language models based, among other things, on the transformer architecture, which learn high-dimensional representations based on very large amounts of language data, and whose performance in numerous concrete language technology tasks far exceeds the performance of purely statistical methods. Accordingly, almost all modern language technology systems use neural methods or large language models in various forms, often also in combination with new symbolic, functional methods (e.g., knowledge graphs with active ontologies), if knowledge is to be represented deterministically, e.g., to also plausibilize results of neural systems (hybrid AI).

Therefore, current research includes combining symbolic and sub-symbolic methods to combine their respective advantages and to compensate for disadvantages, such as integrating complex ontologies or simple knowledge graphs into large language models so that the explicitly encoded symbolic knowledge can be learned from the language model. Corresponding prototypes and language resources as well as commercial solutions and technologies are being developed by about 800 university research groups and independent institutions, as well as about 800 to 1000 companies in Europe. The commercial products in particular are either integrated into existing systems or made available via remote APIs, so that they can in principle be used in any hardware.
In addition to the extremely large amounts of language data required to train language models, very powerful computing systems are needed for this purpose – both aspects represent bottlenecks for actors who do not have access to these resources.
The recommendations presented below come from academic and industrial practice, among others, which the participants of the “Language Technology” group have gained in their everyday work. In addition, further needs and ideas for standardization were collected in various focus groups organized with the consortia of various BMWK (German Federal Ministry for Economic Affairs and Climate Action) and European Union (EU) projects, among others.
In addition to the natural language speech technologies considered here, communication technologies have also developed, such as machine-to-machine communication, which are not considered here.

State of the art in science and technology
General artificial intelligence represents a new generation of AI that can solve tasks for which it has not been specifically trained The goal of development in the field of generalizing AI is to emulate human thinking in its dynamics and diversity, although this does not currently seem foreseeable. Advanced examples in Europe which focus on neural networks include large AI language models such as Generative Pretrained

Transformer (GPT-3) (OpenaAI) or Luminous (Aleph Alpha). Hybrid approaches to general artificial intelligence that have a focus on symbolic, functional knowledge processing based on active ontologies are represented by OntoBroker (semafora systems), for example. Neural language models are trained once with huge amounts of data, causing them to strive for contextual knowledge of the world. They are able to understand and produce a wide range of texts with little human input to support a variety of information-based operations. Symbolic, functional language models extract language deterministically, e.g., from texts relevant for the application, in order to make the knowledge contained therein accessible in a structured way without loss.

In academic as well as industrial research and development (R&D), the focus is mainly on the topics of explainability/transparency, scalability, and metrics for evaluating the sometimes extensive amounts of data required, in addition to application-specific issues. The need to evaluate methods of language processing in terms of their quality is not only very great from a scientific and industrial as well as a user point of view, but also is of societal interest. However, the metrics and methods currently being developed in research are themselves subject to change, since on the one hand the number of individual steps in a testable language processing chain is increasing, and on the other hand the meaningfulness of the metrics used depends to a large extent on the test data (benchmarks) specified. The latter are available for some applications, but are of very different quality.

However, the previously listed challenges definitely affect the available methods differently, e.g. neural methods are strong in scalability and flexibility and symbolic/semantic methods are strong in transparency. Research and industry are therefore working on combining the advantages of both methodological directions in hybrid systems.

State of standardization

In the area of language technologies, there are a number of standardization activities being carried out, which are presented in Annex 13.2. For artificial intelligence, a Working Group and Natural Language Processing for AI is working at European level within CEN/CENELEC JTC 21, which is recommended as a body for the developed recommendations for action

State of regulation

The EU Artificial Intelligence Act (AI Act), which is currently being drafted, addresses all applications that use artificial intelligence and are subject to risk, particularly high risk. For example, among the applications considered high risk, see Draft AI Act; Annex III, are those involving "remote biometrics". AI with speech, especially when speech is used biometrically, can then fall under the regulation of the AI Act and then requires appropriate certification. The AI Act and its requirements relevant to standardization are presented in Chapter 1.4. The EU Commission's draft on standardization topics is also available and should be examined with regard to language technology requirements.

4.1.2.6 Imaging sensor technology

Under the topic "imaging sensor technology", this chapter bundles all AI applications that deal with data from spatial resolution sensor technology. This is intended to cover not only single camera images and image sequences on the visible spectrum, but also other spectral ranges (for example, near and far infrared) or other sensor principles that are imaging in a narrower or broader sense and for which related AI methods are used – such as laser scanners, radar signals, or medical tomography. For the sake of simplicity, this chapter speaks uniformly of "image data", "individual images" and "image sequences".

The associated AI methods, in turn, can be broadly distinguished into three categories:

→ methods that process individual images or image sequences into more abstract information (for example, through object detection or segmentation),
→ methods that synthesize (i.e., artificially generate) realistic image data from abstract, parametric input data,
→ methods that convert image data into other image data of comparable abstraction, for example by „style transfers" from summer to winter photographs.

Thus, imaging sensor technology has great relevance for heterogeneous application fields of high-risk AI (especially in the sense of the EU-AI Act [4], cf. also Chapter 4.3), for example in the fields of automated driving, medical technology or civil security, such as person and face recognition. The planned EU AI Act addresses in particular image processing in the sense of biometric identification as well as high-risk applications in Annex 3. These are the basis for the AI Act to be applied directly or to have an effect via the harmonized standards planned for it, including conformity assessment, and for corresponding certification to be applied. Accordingly, more in-depth descriptions and needs for action on this topic can be found in particular in Chapters 4.6 and 4.7 and as well as in Chapter 1.4. The following chapter only summarizes overarching challenges in these areas in a fundamental way.

In the aforementioned AI areas, modern deep learning methods are, for the first time, achieving results of complex image interpretation that surpass even human capabilities [84]. This has implications for the potentials of the application as well as for risks in the human evaluation of corresponding systems.

Status quo

Existing basic standards for handling corresponding sensor data concern the raw data format (for example JPEG image compression, the Digital Imaging and Communications in Medicine (DICOM) format for robotic sensor data as a de facto standard) and the format of annotations for machine learning, for example the ASAM standard OpenLABEL.

However, the application of machine learning techniques in particular, specifically in the context of deep learning, presents complex challenges On the one hand, these lie in the properties of the AI systems and their development; on the other hand, however, due to the direct and comprehensive data reference of corresponding ML procedures, these challenges also lie in the required data volumes and the associated urgent data protection issues. The primary reason for these challenges is that immense amounts of data are required in the given application field for parameterization and evaluation of AI methods. For example, in addition to a high number of required training examples (as is common with ML methods), there is also a considerable amount of data for a single example image, usually consisting of thousands to several million pixels and multiple colour or information channels. Depending on the application, even only image sequences of several images result in a single training example. This means that imaging sensor technology requires immense amounts of raw data, which – depending on the application – also includes comparably extensive annotations (labels) for training and testing. While there are applications that, for example, assign an entire frame or sequence of images to only a single category or class, there are also applications that require pixel-level annotation of object classes (e.g., segmentation tasks).

Major challenge areas, which are accordingly associated with the application, will be briefly discussed below.

Challenge area "Access to perception data with sufficient data quality"

The provision of the required volumes of training, test and validation data encounters different technical challenges. In particular, processing high-resolution data often requires significant amounts of corresponding high-resolution annotations as a training basis. This is accompanied not only by considerable financial outlay, but also by the challenge of ensuring the quality of an appropriate dataset for specific applications (classification, segmentation, etc.). For high-risk AI applications, for example, the proposed AI Act [4] requires "sufficient relevance as well as representativeness, accuracy, and completeness with respect to the intended application." Evaluating these targets for a dataset consisting of, say, thousands or hundreds of thousands of individual images poses significant challenges to developers and testing facilities. There are already first directories of special cases for the application of image processing such as CV-HAZOP [85]. Human annotation errors here can have far-reaching consequences that are difficult to detect. However, the alternative practice of using AI-based tools rather than human annotation for training and testing purposes also raises questions of verification. Equally complex is verifying that a raw dataset of, for example, street scenes is "representative" or "complete," in particular free of unacceptable bias.

These challenges multiply in AI applications that collect sensor data in an "open world," such as in automated driving (see also Chapter 4.6), rather than in a controlled environment such as is often found in medical or building component testing environments. Here, sensor data exhibit considerable variability, the correct mapping of which in training and test datasets can be safety-critical. For example, the prominent automotive datasets KITTI [86] and Cityscapes [87] do not yet include e-scooters. Furthermore, perceptual datasets are available in varying amounts depending on the use case. For example, 60 perceptual datasets could be found for the automotive sector and only two perceptual datasets for the railway sector [88].

In the environment of automated driving, this challenge can also be specifically seen in standardization, in the transition of the ISO 26262 series [455] (functional safety of road vehicles) and DIN EN 50657:2017 [89] (software for rail vehicles) to ISO 21448:2022 [90] (safety of the intended functionality). This corresponds to an essential change of perspective away from the consideration of primarily stochastic "failures" of a component, towards the consideration of the robustness of an overall system in its environment against potentially also unrecognized challenges (so-called unknown uncertain states and thus "unknown unknowns"), which are essentially associated with the open world. This is seen in the ASAM standard OpenODD (Operational Design Domain), the aim of which is to make permissible application fields for a driving function as precisely specifiable as possible.

Correspondingly advanced considerations in the standardization environment have so far been lacking in other sectors, for example for AI-based construction machinery, human-robot collaboration or civil security. For example, existing standardization activities on AI methods such as ISO/IEC TR 24029-1:2021 [91] and ISO/IEC 24029-2 [92] (which also include an overview of methods) suggest the use of formal verification approaches that are difficult to practically apply to the field of imaging sensing given the scale and complexity of these data. Therefore, empirical test procedures often have to form the basis for robustness analyses (e.g. common corruptions & adversarial attacks). Standardization can make a significant contribution here by establishing guidelines for industrial applications that address, among other things, the following questions in the area of robustness analysis with empirical test procedures:

→ What is the „optimal" testing strategy/process with empirical testing procedures?
→ How can different test procedures be aggregated?
→ How can a risk estimate be extracted from empirical test results?
→ How is a „diverse set" of test methods developed?
→ How do you define „success" in an adversarial attack on an AI model?

Challenge area "synthesis"
The synthesis of image data, i.e. their artificial generation, can, among other things, make a significant contribution to meeting existing data requirements. At the same time, it raises its own challenges and questions depending on the purpose and technology of synthesis.

Synthesis can be used to provide test, training, and validation data for AI procedures. Here, for example, classical computer graphics methods can be used to generate realistic image data. A key factor here is that with appropriately synthesized data, the annotation (i.e., for example, the objects contained in the image and their positions) is usually also directly available, thus eliminating the need for annotation. Raw data of infrequent or high-risk events can also be generated simulatively. But AI methods, specifically machine learning methods, can also be used for synthesis purposes. A significant breakthrough lies in the development of GANs as a machine learning principle that can, for example, generate photorealistic human faces that are not readily distinguishable from real photographs even by humans. However, if synthetic data of an arbitrary generation method are used for training and testing AI methods, essential questions also arise here concerning representativeness and correctness, but especially also concerning the degree of realism of the synthesized data, which has to be proven especially for high-risk applications. Again, there is a significant need, analogous to the need for quality assessment of training datasets, to standardize specific criteria of quality assessment of synthetically generated data or synthesis methods with the specific challenges.

In addition, synthetic data can also be used for non-AI applications, such as art or entertainment However, the methods used for this purpose, especially GANs, can often be used for abusive purposes with little effort, for example in the context of "deep fakes", in which photos of people can be deceptively inserted into video recordings of other people, even realistically depicting their facial expressions and the lighting of the scene. This challenge only concerns standardization to a comparatively small extent, but primarily requires an increase in social competence in dealing with image data. However, a relevant perspective is that authentication procedures in business processes, for example, which have so far been based on photo or video data, should also take this new development into account – for example, through standardized guidelines on which verification or multi-factor authentication steps can be used to exclude corresponding deep fakes.

4.1.3 Standardization needs

4.1.3.1 General

Need 01-01: Cross-sectoral standardization of terms
Especially due to the cross-sectoral meaning of "AI" as a technology, the differences in meaning named above often lead to considerable misunderstandings in interdisciplinary discussions. This creates friction even without substantive dissent and, accordingly, without substantive progress. As the operationalization of AI and AI discussions increasingly require cross-sector and cross-domain measures, it is expected that common terminologies for these will provide a necessary foundation.

As the Glossary makes clear, common terms (for example, "bias," "safety") sometimes have significant variations in standards and conventions across different domains or sectors. It is proposed to create unified definitions across sectors to ensure an overarching terminology especially in AI debates.

Need 01-02: Applicability of the ISO/IEC 5259 series of standards [39] for sector-specific data quality management
Using the ISO/IEC 5259 series [39] as a common starting point for vertical standardization activities in the field of data quality will make it possible to draw on a common framework and to describe terminology, concepts and processes for data quality management across sectors.

With the initiation of the ISO/IEC 5259 series of standards [39], the topics of data quality and data management are being addressed in international standardization, at least in general terms. Nevertheless, it is to be expected that for specific sectors and applications more stringent and possibly different quality criteria than those mentioned above will become relevant. Quality management processes must also be implemented and, if necessary, enriched on a sector-specific basis. Thus, it is recommended to examine, in vertical standardization of data quality management, to what extent the ISO/IEC 5259 series [39] can be used as a general reference and to what extent sector-specific adaptations will be necessary.

Need 01-03: Drawing up of a technology roadmap for AI
A technology roadmap for AI developments can provide a valuable basis for contrasting standardization needs with a detailed timeline of technological developments and needs, and thus sharpening the focus of this Standardization Roadmap in this respect While AI developments are generally very dynamic, trends in processes are foreseeable early on, especially where they play into critical product areas, particularly with regard to ethical trade-offs in AI use (example: the use of neural networks for perception in automated driving). Corresponding developments can be estimated with reasonable robustness and at first independently of concrete standardization needs; at the same time, such a representation can help to identify standardization needs more sharply and along market and technology lines.

In addition to the AI classification methodology outlined in Chapter 4.1.1.1, it is recommended that support be given to work on the development of a technology roadmap that summarizes current technology trends in AI and makes recommendations for the future development of Germany as an industry location.

Need 01-04: Test standard for AI systems following the CC
Since the CC is a globally accepted approach to security evaluation of IT systems used by testing laboratories and certification bodies, this avoids or minimizes additional effort in product certification of AI systems by relying on best practices.

For the testing and evaluation of AI systems, a horizontal testing standard is to be developed that is based on the Common Criteria documents in terms of terminology, methodology, and structural specifications.

Need 01-05: Requirements for certification bodies
Required auditor competencies and the time required for an audit according to ISO/IEC 42001 [27] may differ from audit requirements in other areas.

Formulation of requirements for certification according to ISO/IEC 42001 [27], which must be fulfilled by certification bodies. A German project proposal on this topic is in preparation; however, the project implementation must be significantly supported by the German side.

Need 01-06: Standardized form of describing AI solutions
Complementary to the draft EU AI Act, a standardized form of describing AI solutions should be made available. Chapter 4.1.1.1 of the German Standardization Roadmap AI presented here can be used as the basis for this. The corresponding test methods can also be precisely detailed on the basis of such a scientifically sound description of the AI technologies used. Regulatory and certification efforts would therefore decrease, while quality would increase. The same applies to the description of entire AI applications in which, for example, several AI technologies are used:

→ This could also significantly improve the required technical documentation (Draft AI Act Art. 11), increasing the transparency (Draft AI Act Art. 13) and trustworthiness of AI.
→ The European AI Database for managing "high-risk" applications listed in the EU would also benefit from a unified taxonomy for describing AI.
→ Furthermore, it is conceivable that, using the proposed AI classification, uniform "harmonized European labels" will emerge in the future to further promote and accelerate the dissemination of transparency and quality of AI.
→ Another important point is that conformity assessments would become simpler and easier to standardize through uniform classes of AI applications. The same applies to market surveillance.

It is therefore recommended to initiate a standardization project for the classification of AI systems on a European level.

4.1.3.2 Ethics

Need 01-07: Design interfaces for the AI development process
Standardized interfaces and a modular model of typical AI building blocks can enable interchangeable development and individual evaluation according to standardized criteria, thus contributing to overarching usability, transferability of approvals, and transparency. Appropriate methods for viewing models and datasets are also required by the draft AI Act [4]. Building on this, standardized process models can be created (cf. [93] for example) that integrate the provision of appropriate interfaces as a regular artefact of development and minimize additional effort. The resulting comparability of the interface management of different institutions will provide orientation and thus contribute to the value of self-determination in the sense of self-determined use.

Standardized interfaces in AI systems should provide external auditors with insight into, for example, training datasets and models already in the development phase, and should merge AI subsystems, where possible, into common uniform functional descriptions in order to simplify development, testing, and deployment, especially with regard to ethics and trustworthiness goals (for example, with regard to traceability, authenticity of data, transparency). Standardized role descriptions of AI components and actors should be defined. Furthermore, a standardized description of the interaction of the individual components among each other as well as in the overall context (including non-AI system parts and system environment) should be created. Which degree of abstraction is advisable in practice is to be defined – for example, in order not to have to disclose all components of a dataset, but only abstracted characteristics, in consideration of data protection, data economy and data scope.

Need 01-08: Design of the contents of a quality backward chain
In order to also be able to evaluate artificial intelligence systems in their ethical dimension during their use and, if necessary, to model decision bases, the use of a quality backward chain is recommended. This acquires field data as part of the deployment, which enables a judgement to be made about ethical decisions made by the system. Fundamental corrections to the system are not envisaged here; rather, the aim is to ensure the ability to respond (appropriately) to damage caused by deployment. The quality backward chain provides data for the subsequent assessment of possible wrong decisions and thus helps both the provider and the user.

Mandatory content in the context of field data collection in the sense of a quality backward chain, which must systematically cover not only technical but also ethical aspects, requires standardization as well as uniform data formats to ensure future reporting obligations. This is to ensure that the option to make reports is as low-threshold as possible and is as accessible as possible for all user groups. This should ensure democratic use with regard to the value level. This is also necessary with regard to interoperability in order to enable the free use of products, services and systems away from monopolies, and to support users in their sovereign decision-making in this respect as well.

Need 01-09: Provide opportunities for re-evaluation
The ethical re-evaluation of AI systems is based on their core values. These core values must be identified beforehand in the development process by the company as part of a stakeholder process. On the basis of the values that have been considered, the company internally classifies the results or decisions of the AI system in terms of its ethical dimension in operation, but also already as part of the development process. Field data from a quality backward chain can support this evaluation. The relevant stakeholders must be involved in the review. It can be completed by a panel of experts, such as an Expert Review Board, or other trained personnel. The review also includes looking at the company's processes with regard to ensuring ethical principles and correcting them if necessary. If a violation of the above-mentioned core values is discovered, a major review of the processes and data basis will be required. A reporting obligation analogous to that in the case of data protection violations would also be conceivable. Re-evaluation should take place on an as-needed basis or at fixed intervals, for example every three years. Core elements of this process are already addressed in ISO/IEC 38507:2022 [26], where for the most part the core objectives of the company are placed in the foreground and ethical aspects occur more as a secondary requirement. Moreover, it has not been worked out which concrete contents are to be taken into account with regard to the ethical evaluation and to what extent they are to be considered.

Documentation requirements and intervals for mandatory re-evaluations are to be standardized.

Need 01-10: Standardization of a concept for privacy ethical design
Privacy ethical design underpins all systems with the principle of individual privacy. In doing so, it goes beyond the concept of privacy per se and assigns it a clear ethical dimension, taking into account not only direct influences but also indirect influences on the needs of the user. This promotes a basic trust in new technologies and thus increases market acceptance. Interoperability between different providers, such as SSO, can also appeal to more users through privacy ethical design. This can be done taking into account the project currently initiated in ISO/IEC JTC 1/SC 42 on an MSS for AI (see Chapter 4.1.3, Need 1 "Support of international standardization work on an MSS for AI"), by including the explainability of AI systems in the list of requirements of the emerging document, and by extending the concept of risk to include ethical risks, as has already been done in the ISO/IEC 23894:2022 [25] Risk Management project.

To promote effective privacy ethical design, ethical risks must be specifically and systematically addressed. As part of a risk management process, they are to be identified and analyzed in order to mitigate them through targeted measures. This can be designed, for example, in the form and scope of a possible documentation requirement – to promote transparency and prevent purely apparent measures. Such an approach contributes to the value of traceability, among other things. Another example would be to improve the user interface with regard to privacy settings in order to create the best possible opportunities for stakeholders to implement privacy effectively and intuitively.

Need 01-11: Design purpose limitation of data
In order to enable all parties to act transparently in the interest of trustworthy AI development, it is necessary to further develop the purpose limitation of data. According to Article 5 of the General Data Protection Regulation (GDPR), personal data shall be "collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes". Exceptions to this are given in Article 89 of the GDPR for "processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes". This is where standardization can come in and, within the legal guidelines of the GDPR, promote innovative data use that enables companies to develop new products based on their master data without violating consumers' rights. A good option on this point is to maintain a dialogue on relevant legislative proposals of the European Commission (DSA, DGA) in order to further the regulatory intentions in this respect in a coherent manner. At best, the expertise of the supervisory authorities should also be included in this process via joint exchanges. At the same time, consumers should always be able to obtain an appropriate overview of the purposes for which their data is used. Standardization can support companies and institutions in developing and integrating the necessary consent management.

For the secure and innovative purpose limitation of data, standardization can promote uniform documentation and consent forms that provide users and providers with quick and uncomplicated insight into the possible uses.

Need 01-12: Design of the value system
This need assumes that currently no ethical behaviour can be assigned to machines; however, the technological implementation implicitly or explicitly allows conclusions to be drawn about ethical assumptions in the development process. These may include, for example, the mechanisms by which fairness or safety (in terms of risk considerations) are implemented. It should be possible to gradate the degree to which defined values are implemented via AI depending on the use case, combined with justifying documentation for this decision. Here it is recommended to work towards a standardization of the representations, which reduces development risks and creates social transparency in the organization. This standardization is to be oriented to the relevant standards-setting and the social discourse on the phenomenon concerned. In addition, this could be integrated into the guidelines of companies, for example the Code of Conduct, binding for all employees (cf. ISO/IEC 38507:2022 [26], ISO/IEC 42001 [27], the latter currently under development). There are numerous phenomena in which the question of a value system of AI plays a role. One example is the question of the extent to which AI can be used in recruiting processes to systematize and categorize job application data. At this point, the aspects of diversity and fairness are involved, among others. Another illustrative example is the question of whether risk assessments in operation (such as in the case of risk-based planning algorithms in automated driving or dynamic risk management [94], [95]) can be permissible compared with the stipulations of the Ethics Commission for Automated and Connected Driving of the Federal Ministry for Digital and Transport (BMVI) [96]), and if so, what concrete requirements can be derived from this for technology. It is estimated that this mapping of ethical concepts into machine-readable form cannot yet be implemented directly in standardization, but requires focused interdisciplinary preliminary work by

research (see also research fields artificial ethics and artificial morality).

Where ethical concepts are to influence the decisions of AI systems at run-time, their formalization and representation in machine-usable form is required, for example in the form of ontologies or in the form of computational principles for permissible risk considerations.

Need 01-13: Improved and lower-threshold overview of the interplay between criticality levels and associated requirements (especially for low-risk AI systems)
In order to be able to quickly classify AI systems in terms of their criticality and to be able to grasp the associated requirements well, clearly structured provisions would be helpful for manufacturers. This is especially true for the question of what requirements low-risk AI applications should meet in order to comply with regulatory requirements, but also to achieve a high level of trustworthiness. The planned AI Act does provide a classification into certain classes for the EU area, e.g. by defining prohibited areas or also high-risk systems, whereby the classification is primarily made according to the area of application and less according to the risk arising for the respective concrete product. However, few concrete requirements remain, especially for the area of less critical systems, so that in this case the manufacturers do not get a clear picture of which requirements are to be implemented. This effect is reinforced by the fact that there is now a wide variety of other laws at EU level, such as the General Data Protection Regulation, the Digital Service Act, the planned Data Act, and also the EU Charter of Fundamental Rights, among others, which provide further important requirements that play a central role in the development of AI-based systems. This also makes it confusing for users as to how they should classify the systems, what constitutes a trustworthy system, and which requirements they fulfil in what way.

Better transparency and clarity with regard to the different levels of criticality (also beyond the classification in the planned AI Act [4] and the associated requirements) should be created and anchored in corresponding standards. The aim is to convey in a low-threshold manner what constitutes trustworthy AI, how the systems are to be classified, and which requirements from which laws are to be implemented.

Specifically, this includes the following points:

→ A low-threshold classification of AI applications that is transparent for manufacturers and users, with regard to the criticality of such applications.

→ For manufacturers: a targeted clarification of which requirements from which legislation are to be implemented for which applications or criticality levels in order to be able to develop legally compliant and trustworthy AI systems. Suitable standards/tools should be used to create an easily understandable overview that breaks down the relationships between the requirements, the associated laws, and the steps required for each use case.

→ For users: a quick and low-threshold insight into the different levels of criticality and their requirements at an understandable level in order to make the trustworthiness of AI systems ascertainable in an appropriate manner.

4.1.3.3 Quantum AI

Need 01-14: Artificial intelligence (especially machine learning) and quantum computing in the context of IT security
The use of quantum computing has the potential to have a profound impact on the practice of artificial intelligence, particularly machine learning. Although the state of development of current quantum hardware still imposes strong limitations on the current application of quantum AI to practical problems, significant progress can be expected in the next few years. Numerous national and international funding projects are also being driven forward with commitment with regard to the corresponding quantum software stack. Quantum AI methods play a very crucial role at this point as components of the first essential applications for quantum computers and especially already in the NISQ era.

Developments in the field of quantum AI, and in particular in the field of quantum machine learning (QML), will need to be continuously monitored and thoroughly evaluated in the coming years. This presents a great opportunity to work towards the secure design of the new technology at an early stage and to identify and address both potentials and risks in its use. Looking at the current state of the art, this requires targeted research efforts and related activities that, on the one hand, investigate and strengthen the security properties of QKI systems and, on the other hand, consider the use of QKI within IT security. The realities and challenges of the corresponding quantum infrastructure must also be considered beyond the mere focus of QKI models. These include the interfaces between classical IT and quantum computers in the form of hybrid systems, as well as the fact that the dissemination and distribution of quantum computers will in all likelihood not be the same as that of classical IT due to their technical nature. Overall, a consistent pursuit of the aspects just mentioned is essential to enable the development of suitable security standards for QML systems in perspective. Synergy effects between the multitude of nationally and internationally funded projects for the development of quantum hardware and software, as well as the respective security requirements can only be exploited if early coordination and exchange processes take place for this purpose.

4.1.3.4 Language technologies

Need 01-15: Standardization of language technology and natural language processing APIs and data structures
The APIs of language technology cloud services are not standardized and are therefore different in each case, which makes the comparison, testing, benchmarking and exchange of different APIs difficult or impossible, i.e. there is currently no interoperability. For in the best case automated usability of data collections, it is necessary to standardize metadata descriptions in such a way that all essential properties of a data collection are available in machine-readable, semantically annotated form. Numerous initiatives are working on this issue, in particular the National Research Data Infrastructure (NFDI), European Open Science Cloud (EOSC) and Gaia-X.

For automatic speech recognition (ASR) processes, there are also no specifications or guidelines as to how, for example, punctuation or numbers should be handled, i.e. transcribed. Standardization is necessary for better comparison, for benchmarking and also for the exchange of corresponding services.

The DFKI has already gained initial experience in this area within the framework of the EU project European Language Grid, as well as presenting initial proposals with the cooperation of the University of Sheffield. This aspect also affects a number of associated topics, e.g. annotation formats, workflows, benchmarks, transfer learning for language models. The problem is: All vendors each follow their own philosophy, i.e. they offer different, proprietary APIs. It would be helpful to be able to evaluate and thus compare a vendor's technologies using standard datasets (or their own data) and standard metrics (e.g., WER for ASR). This topic also concerns large language models, i.e., in particular, how language models are addressed to perform transfer learning. Of relevance for industry: No company can develop a large language model on its own, so fine-tuning and transfer based on standardized methods and interfaces are mission critical to adapt the language model to the specific use case.

At least a Europe-wide standardization of **language technology and natural language processing APIs** with regard to functional scope and parameterization should take place, so that more interoperability and also better comparability between the cloud services of individual providers is created. In this context, data formats, e.g. regarding data exchange, and semantic annotation formats can also be considered. This includes standardization of **metadata**, data collections, data sheets, model cards, language models, accessibility, use of data and data collections for research and commercial applications (can be embedded in NFDI, EOSC, Gaia-X, etc. as appropriate). Furthermore, it is helpful to standardize **guidelines for transcription** procedures that often include ASR or build on ASR output, e.g., number as number, number as word, etc., punctuation, capitalization, etc.

This point also includes the **orchestration** of services in the form of **workflows** or **pipelines**. This aspect also concerns the standardization of **benchmarks** for comparing diverse solutions, e.g. ASR or natural language understanding (NLU). In the context of dialogue management applications, this aspect also concerns the standardization of **resources for modelling dialogues.**

Need 01-16: Standardization of the measurement of performance, correctness, precision and plausibility of large language models as well as data quality
Language models currently represent the state of the art in research and technology for many language technology applications, but there are still no standards for or measurement of fundamental properties such as correctness, precision, facticity, self-consistency, etc. – among others, to be able to assess a language model and compare different language models. Self-consistency of a model may include, for example, whether a model contradicts itself on certain related questions. (Note: However, rule-based/symbolic models are already part of hybrid systems or pipelines today). For example, measuring the degree of truth of the output of language model-based applications (or the self-consistency of the model) – if technically feasible (and if only in some well-defined domains) and resiliently realizable – can signal the quality of the language model. It should be noted that increasingly multimodal models, image understanding, combination of language and images, sign language (recognition and generation) are also performed on the basis of large language models (Stanford also calls these foundation models).

Text, audio, and video data, among others, are used to train language models and other machine learning techniques. Currently, there are no standards for measuring the quality of such data and data collections, including for determining whether they should be made usable as part of a training dataset. Standardized methods for measuring data quality are also highly relevant to the aspect of data bias.

Standardization of the **measurement of performance, correctness, precision, plausibility** in the respective application context of large language models is desirable. In this context, measuring the quality of the output of generating language models is also relevant, e.g., in terms of meaningfulness, grammaticality, semantics. Here, there is a need for standardized metrics. Furthermore, the term "language model" would have to be defined, with respect to a differentiation from text-processing, possibly also rule-based models.

The standardization of approaches to **measure data quality for language models**, i.e., text quality in particular, but also audio quality and video quality, are relevant for compiling datasets used e.g., for training language models, as well as for measuring bias. This concerns, among other things, the selection of data to be used for training language models, e.g., to assess/avoid bias and hate speech, etc. Approaches for describing and measuring bias itself (including a specification of the different dimensions of bias, e.g. political bias, gender bias, etc.) also need to be described and standardized.

Need 01-17: Knowledge graphs and ontologies in large language models
While language models represent the state of the art in science and technology for a variety of language technology tasks, there are numerous knowledge bases, knowledge graphs, and ontologies that contain symbolic knowledge or semantic knowledge in symbolic representation. Currently, there are no standards on how such knowledge bases and ontologies can be integrated into language models and made usable in a **secure** way (evaluation of "criticality") according to the respective requirement. This aspect also concerns the merging and integration of different knowledge bases and knowledge packages.

Standardizing approaches, such as **knowledge graphs and ontologies in large language models** that can be integrated and made usable, serves to leverage existing symbolic knowledge assets in the context of state-of-the-art language technologies, which are typically based on large language models. This should also consider the merging, integration and management of ontologies and ontology modules or ontology packages from different sources. This can also consider approaches to integrating (more ontology-based) world knowledge into (more document-based) knowledge graphs. These aspects are important and relevant for the use of symbolic knowledge bases (i.e., ontologies) in the context of knowledge graph-based applications.

Need 01-18: Testing and auditing processes for AI language applications
In the context of trustworthy AI, standardization of **testing and auditing processes** will also gain importance for (learning and continual learning) NLP systems.

In particular, when NLP systems such as search engines, recommendation systems, or chatbots serve as decision support systems in critical applications, it will be necessary to define testing and auditing processes. In addition to the direct variables (type and generation of test items, metrics for evaluating the results), this also includes the question of process participants. For example, in healthcare, it may be appropriate to involve patient representatives in a participatory process in the design and execution of testing. Continual learning systems will need to be retested and audited in specific cycles. Here it is necessary to define the criteria according to which the cycles are determined.

Need 01-19: Supporting digital language equity
Of the numerous European languages, only some are well or very well supported by technologies. In addition to English, these include French, Spanish and German. To measure and assess the support of a language by language technologies, recent results from the EU project European Language Equality are available: the Digital Language Equality Metric. Such a metric could be standardized across Europe so that respective language communities can set language-specific targets and key performance indicators (KPIs) in the context of all languages in Europe, which can also be monitored collectively.

Digital language equity – ensuring that all languages in a language community (city, region, organization, etc.) of whatever scale are supported by language technologies in a similar, balanced, equitable way – is important for the internationalization of content and technologies, and for the scalability of technologies.

Need 01-20: Reviewing standardization requirements from the proposed AI Act for language applications and adding to them as necessary
Language can potentially fall under the regulation of the AI Act as a high-risk system and be subject to its requirements such as risk assessment, quality management or demonstration of robustness, transparency. For the standardization requirements from the planned AI Act, investigations of existing standards are required for language technologies in order to clarify to what extent the requirements are already covered and what may still need to be supplemented with standards. Research and development activities may be required to provide the desired methods, e.g., for "record keeping through built-in logging capabilities" or for "robustness specifications."

The aforementioned standardization requirements from the planned AI Act require adapted standardization, e.g., as biometric systems of an AI, although the requirements may not have been fully technically researched and developed. Further standardization and research is recommended for this purpose.

4.1.3.5 Imaging sensor technology

Need 01-21: Development of evaluation metrics and methods for image datasets and collection/synthesis procedures and image-evaluating ML procedures
Datasets are increasingly taking on the role of parameters, especially in modern ML methods. Accordingly, requirements for AI systems are also being formulated by means of requirements for datasets, for example in the draft AI Act. However, there is currently a lack of standardized procedures that could be used to describe quality properties of datasets across the board. However, case-specific procedures do not achieve any comparability and thus limit the assessability of different AI procedures. A standardization of corresponding methods for quality assessment as well as the targeted development of standardizable, cross-application methods can contribute significantly to a better, overarching understanding here – even if the standardized metrics do not fulfil the claim of an undisputed, absolute quality criterion, but merely enable a transparent, overarching indication.

Standardized evaluation metrics should be developed that allow either datasets (from real or from synthetically generated image data) or methods that generate these datasets to be evaluated according to common quality criteria. These metrics should address common targets, for example under the draft EU-AI Act (cf. [4]), "relevance, representativeness, freedom from error, and completeness with respect to the intended application") and specify appropriate measurement principles of these targets. These metrics should be largely independent of AI methods or applications, but clearly identify limiting assumptions/applicabilities where necessary. Existing approaches (for example [97]) are to be examined for suitability. Where no suitable procedures exist that can provide an estimate, new approaches are to be developed within the framework of R&D.

Need 01-22: Standardizing metrics for testing image-processing AI systems
Analogous to the need to "develop evaluation metrics and methods for image datasets and collection/synthesis methods," there is a need to standardize metrics that enable the evaluation of image-processing AI systems while defining their application domain. For example, the metric of "mean intersection over union" (mIoU) has been established in the scientific community for evaluating ML methods for image segmentation. Corresponding metrics should also be provided for other tasks such as object detection, classification, or image conversion. Jointly standardizing analogue metrics for common AI methods can contribute to the comparability of heterogeneous approaches.

It should be noted that the metrics may contain risk-dependent components (e.g., risk-dependent assessments of segmentation errors, e.g., in critical regions for medical imaging data). In doing so, these mechanisms should be designed to be generic/model-agnostic so that they can be easily applied to different scenarios.

Need 01-23: Method for cyber-secure authentication based on image data
Methods must be developed to assess the extent to which given image features are still trustworthy according to the state of the art (and can therefore be used for authentication) and at what point corresponding features can be manipulated, for example by "deep fakes". Procedures for ensuring the authenticity of identities and information must be specified, on the basis of which an appropriately required level of trust can be established for different applications.

Need 01-24: Developing metrics for assessing privacy risks by reverse engineering ML models
ML models can store personal information from the training datasets, for example in the context of "overfitting", so that

this fact is unknown to the developers, but the information can be improperly reconstructed by knowledgeable attackers. Targeted research to assess the risk should work toward perspective standards for the development or assessment of ML models in application that make these risks manageable.

Metrics should be explored to assess what type of personal information may be latent in a given ML model and how to mitigate corresponding potential for misuse.

Need 01-25: Conducting and promoting more research on a data protection-secure development of AI/ML

Despite the often-held assumption that data needs in ML applications are incompatible with data sparsity or anonymization, it should be noted that it is unclear whether there is not a clear potential to reconcile data protection and AI/ML performance without having to accept significant shortcomings – provided that appropriate procedures are made available. For high-risk applications in particular, however, it is necessary to prove that performance is not in fact unacceptably impaired by anonymization, for example. In order to achieve a well-founded assessment and either apply suitable procedures as standard or, if necessary, make a justified trade-off between, for example, safety and privacy, a quantitative assessment of technical potentials and risks based on scientific findings is required.

Targeted research should be conducted and funded with the goal of enabling the development of high-performance AI/ML methods for image data while complying with data protection constraints, and quantifying the extent to which this is possible. This means, for example, the development of metrics to estimate performance losses due to anonymization or data sparsity, as well as the development of suitable anonymization methods, the development of AI/ML methods that are robust to anonymization, or the development of methods that specifically limit data collections to relevant cases, thus reducing data volumes of collection, storage and annotation. Suitable methods are to be standardized in the future.

Need 01-26: Conversion of DIN SPEC 13266:2020 [98] into a standard

It seems there aren't any standards for deep learning systems.

DIN SPEC 13266:2020 [98] is a specification for deep learning systems and describes the current state of the art very well. It should be turned into a full standard.

Need 01-27: Extension of ISO 21448:2022 [90] to other use cases

The main content of ISO 21448:2022 [90] fits most mobility applications. Part of the content can also be used for cases other than mobility.

In the title, ISO 21448:2022 [90] only refers to road vehicles. It should also be extended for other use cases beyond mobility.

The Working Group Basic topics ranked the identified needs according to the urgency of their implementation. Figure 21 shows the urgency of implementation, categorized according to the target groups of standardization, research and policy.

Urgency of implementation →

Need	short-term			medium-term			long-term		
Standardization									
01-01			■						
01-02			■						
01-04			■						
01-05		■							
01-07					■				
01-08					■				
01-09					■				
01-10					■				
01-11					■				
01-13							■		
01-15		■							
01-16		■							
01-17					■				
01-18		■							
01-19					■				
01-20		■							
01-21			■						
01-22				■					
01-23						■			
01-26					■				
01-27								■	
Research									
01-03			■						
01-12						■			
01-24						■			
Policy									
01-06		■							
01-14	■								
01-25			■						

Figure 21: Prioritization of needs for the key topic Basic topics (Source: Working Group Basic topics)

4.2 Security/safety

The fundamental need for testing and certification of the safety, security and privacy properties of an AI system arises almost naturally from the context of the use of AI systems in existing processes and products and the existing requirements for risk minimization and secure/safe operation in that context.

The task of this chapter is to develop recommendations for action that make it possible to use existing testing and certification models from product safety and IT security for AI systems as sensibly as possible. Also, AI systems should be provides with the possibility to increase their security/safety by means of suitable procedures (controls) and with the possibility to demonstrate an appropriate level of security/safety. As already pointed out in the first edition of the Standardization Roadmap, people trust in safety-tested coffee machines or safe components in nuclear power plants, as well as in trained personnel, and for all these areas there are corresponding standards and methods that make the degree of implementation of a standard assessable and thus make the safety gain certifiable. Ultimately, it is precisely this certified proof of compliance with basic principles in all areas of safety and security of an AI system that should help to create trust.

4.2.1 Safety

4.2.1.1 Status quo

In the opinion of the authors, the topic of safety is of particular importance and it was dealt with intensively in the sector-specific chapters in the first edition of the Standards Roadmap, with many cross-sectoral aspects. That is why the topic of safety (product safety) was included as a horizontal topic in this Roadmap. Further sector-specific safety aspects are described in Chapters 4.6 and 4.7.

The term "safety" is closely related to the term "risk," although the term "risk" can be understood in different ways. The German term "Sicherheit" can mean "safety" as well as "security". Therefore, a clarification of terms follows first. Then, the topics of "AI" and "Safety" are related and two types of relationships are derived: "direct safety relationship" and "indirect safety relationship". Then, the two types of relationship are examined in more detail. Finally, a conclusion is drawn with key recommendations for action.

In the following, we first explain how the term "safety" is related to the term "risk" and how it is distinguished from other quality characteristics. The understanding of the term "risk" in the proposed AI Act is then discussed. Finally, we determine how the terms will be used in the rest of the chapter.

"Safety and risk" according to ISO/IEC Guide 51:2014 [99]
ISO/IEC Guide 51:2014 [99] gives guidance for work on standards and guidelines regarding the inclusion of "safety". "Work on standards deals with safety aspects in many different forms across a wide range of technologies […]." In this Guide the term safety stands for "freedom from risk which is not tolerable", risk standing for the "combination of the probability of occurrence of harm and the severity of that harm" and harm is defined as "injury or damage to the health of people, or damage to property or the environment".

The existing chain of consequences in the area of safety (functional safety) is often extended into areas that are not absolutely independent of it, but do not belong causally to the topic of safety. An example would be the failure of a power plant, which is problematic for the availability of the power supply, but is not actually a safety issue. Nevertheless, a power failure can lead to concrete damage for humans as well.

In addition to the consequence, however, there are also extensions of the causal relationships in the direction of the causes, e.g. the subject area of security (information security). It should be noted here that the cause of a manipulative intent on the part of the attacker requires a dimension of consideration that is not originally included in safety risk analyses. The debates about the meaning of the topics of safety and security have been and are still being debated in technical regulation and standardization, whereby there is consensus that both topics must be considered and that security is considered a basic prerequisite for safe operation in the sense of safety. With regard to the cross-sectional topics of safety and security, reference is made to documents such as the technical report DIN CLC IEC/TR 63069:2021 "Industrial-process measurement, control and automation – Framework for functional safety and security" [100], as well as to the work results of the maintenance team on DIN EN 61508-1:2011 [101], DIN EN 61508-2:2011 [102] and DIN EN 61508-3:2011 [103] which have dealt with these issues over a longer period of time, in order to establish that it is an important prerequisite that the safety consideration presupposes effective protection by security measures and can only be valid if this prerequisite is assumed. This chapter takes up this approach and refers later to the area of security in the case of causal security threats.

“Risk” according to the EU AI Act
With regard to AI, the EU is pursuing a “risk-based approach of AI regulation“. Here, the meaning of the term risk is broader than in the context of “safety” or the ISO/IEC Guide 51:2014 [99]. It also refers to risks relating to fundamental rights. This includes issues such as data protection, freedom from discrimination, protection of privacy and protection against subliminal manipulation of individuals, see Chapter 1.4.4.

Discussion and definition of terms
Within the meaning of the EU AI Act, any safety-relevant AI system falls into the “high-risk” category. In terms of ISO/IEC Guide 51:2014 [99], high risk has a different meaning, as risk is fundamentally about safety risks. These safety risks can be marginal and thus acceptable to very high and thus unacceptable. In the following, the terms “safety”, and “risk” are used in the sense of ISO/IEC Guide 51:2014 [99].

Regarding the term “AI system,” the definition of the proposed AI Act is used. This definition clearly states that an “AI system” is a specific type of software. However, it is very broad and fuzzy in terms of defining the type of software. It does not define a clear boundary between conventional software and AI software. In the following, we will elaborate on the role of AI software in the context of safety.

Relationship between AI and safety
Safety is achieved through an iterative process of risk identification, risk assessment and risk reduction. An AI system can play a role in this risk management process in a number of ways. In the following, the risk management process is explained. We then look at how software in general, and AI in particular, can play a role in this risk management process. It then addresses the difference between AI being able to be used to realize the “normal” behaviour of a system or being used to achieve the necessary risk reduction. Based on this, specifics regarding the behaviour of autonomous systems in complex environments are discussed. Finally, two classes of safety references are presented: “AI with direct safety relationship” and “AI with indirect safety relationship”.

Iterative risk management process according to ISO/IEC Guide 51:2014 [99]
The Introduction to ISO/IEC Guide 51:2014 [99] states that “the increasing complexity of products and systems entering the market makes it necessary to place a high priority on consideration of safety aspects”. The section below describes the approach to risk identification and risk assessment, as well as the corresponding reduction of risks over the entire life cycle of the product.

Risk assessment is the first and most important step in planning and evaluating suitable safety measures. ISO/IEC Guide 51:2014 [99] describes the most elementary considerations for this and the steps to be taken to achieve an acceptable level of risk (see diagram in Figure 22).

4.2.1.2 Requirements and challenges

Software (AI or conventional software) in the risk management process
AI as software can play a role in different ways for different steps in the process. For example, software can be the object of consideration in a process step or be used to perform a process step, thereby moving it into the scope of risk assessment.

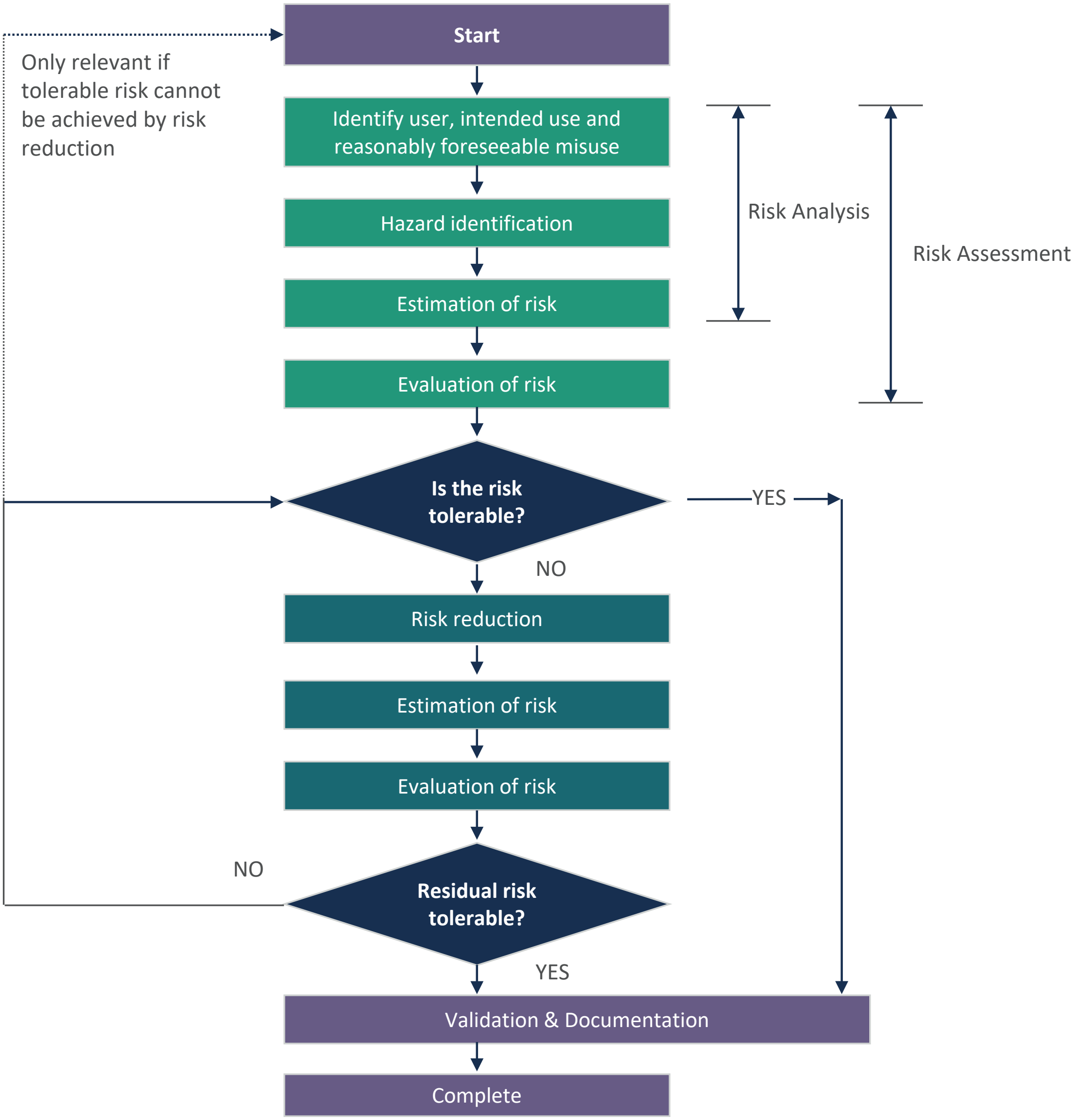
Software itself does not represent a hazard in the sense of safety, but it is decisive and increasingly responsible for the behaviour of technical systems. Software can contribute to the emergence of hazardous situations due to system behaviour. The relationship of software to safety is therefore always about the behaviour of a technical system, and this also applies to AI.

However, the behaviour of a technical system never depends on software alone. It typically results from an interaction of software, hardware and other elements in a possibly complex environment with people and other technical systems. Software must always run on hardware. Accordingly, safety considerations must be applied to software and hardware in combination. This aspect comes up short in the planned AI Act. Hardware faults and failures are to be considered in the AI system robustness requirement, but since the AI Act specifically refers to software in the AI system definition, the hardware relationship is not explicitly addressed.

Nominal behaviour vs. risk reduction behaviour
AI can have an influence on the nominal behaviour of the system in terms of “intended use”, e.g. in the driver assistance system of a car (see Figure 22).

Nominal behaviour can contribute to hazardous situations arising from hazards such as mechanical hazards, electrical hazards, thermal hazards, etc. (see e.g., [517]). No new risk approaches are needed to assess the risks posed by the

Figure 22: Iterative process of risk assessment and risk reduction (Source: along the lines of [99])

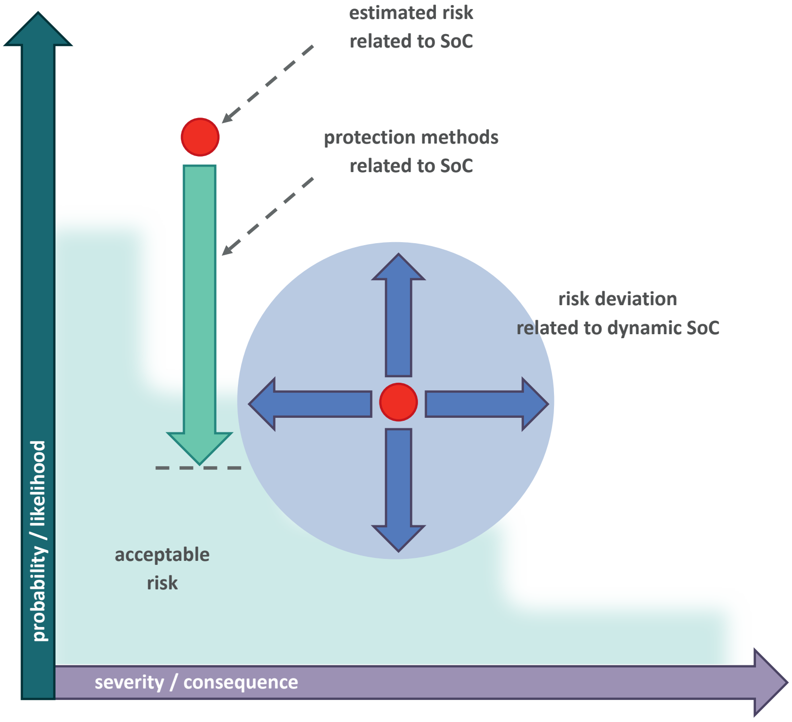


hazardous situation. The evaluation of whether a particular system behaviour is acceptable or too risky does not depend on whether AI or other technology is used to realize the specified behaviour. However, AI is often used to realize functions that are difficult to fully specify. This is especially true for systems that are intended to perform complex tasks in a complex and changing environment automatically, i.e., without intervention by a user. For such complex scenarios, the challenge is to ensure that all relevant situations are considered and taken into account in the specification of the system behaviour. Depending on the type of AI, it may also be that the nominal behaviour changes. These aspects can lead to uncertainty in risk assessment. The uncertainty of the risk of a system under consideration (SoC) is shown in Figure 23 by a blue circle around a red dot. The red dot refers to the risk of the SoC and the blue circle describes how this risk can deviate due to the dynamics of the SoC or its deployment environment.

Figure 23 also describes the case where AI is used to reduce an identified risk. AI could be used in protective mechanisms, for example.

Significant risk reduction can be achieved with protective mechanisms, but the requirements for correctness of implementation are correspondingly stringent. The extent to which AI can meet these requirements for correctness is the subject of current research. Depending on the type of AI, complexity of the protection mechanism, and the deployment environment, the current state of knowledge and technology may or may not be sufficient.

Figure 23: Risk diagram (probability-consequence assessment) (Source: Holger Laible)



Another way to reduce risks is to provide warnings. The use of AI for the realization of warning functions is typically associated with lower risk reduction and thus with less stringent requirements for the correctness of the implementation. Nevertheless, the current state of the art in science and technology in AI may not be sufficient to meet the requirements, and there is a need for research in this area as well.

Autonomous systems in complex environments

With increasing levels of automation, complex tasks, and unstructured environments, it becomes more difficult to produce a specification that is sufficiently complete, free of contradictions, and describes behaviour that is acceptable. Accordingly, there are discussions regarding the basic assumptions on the safety of nominal behaviour. For example, is it safe for a driverless vehicle to pass a parked car at a certain speed (a standard case in normal driving), where theoretically a small child could suddenly jump out? This is a situation that is risky with human driving, but occurs. What should an acceptable operating behaviour look like in such a context and is there any possibility at all to cover all eventualities? It is not possible to know at the time of designing what will happen during operation. As long as what is not known is known, it can be dealt with, but in complex environments there are also unknown gaps in knowledge ("unknown unknowns").

The challenges regarding increasing autonomy and complexity of mission and task were addressed in the paper "Autonomy Levels for Unmanned Systems (ALFUS) Framework, Volume I – Terminology" [104] back in 2007 (see Figure 24).

The three dimensions "mission complexity", "environmental complexity" and "autonomy (human independence)" initially have nothing to do with AI, but with the characteristics of the overall system. However, AI is often an essential tool for trying to manage complexity and achieve a high level of automation. With regard to safety, however, the corresponding tasks and expectations are in conflict with the KISS (Keep It Simple, Stupid) approach to safety solutions.

The safety of the nominal behaviour is particularly relevant in automated driving, where it is partly discussed under the concept "safety of the intended functionality" (SOTIF). However, SOTIF is primarily concerned with finding errors in the behaviour specification, rather than whether a specified behaviour is too risky or still acceptable.

One area of research that deals with making autonomous systems safe is dynamic risk management. The aim is to give autonomous systems the ability to assess and control risks themselves. This does not mean that autonomous systems acquire any real understanding of risk. It is about developing

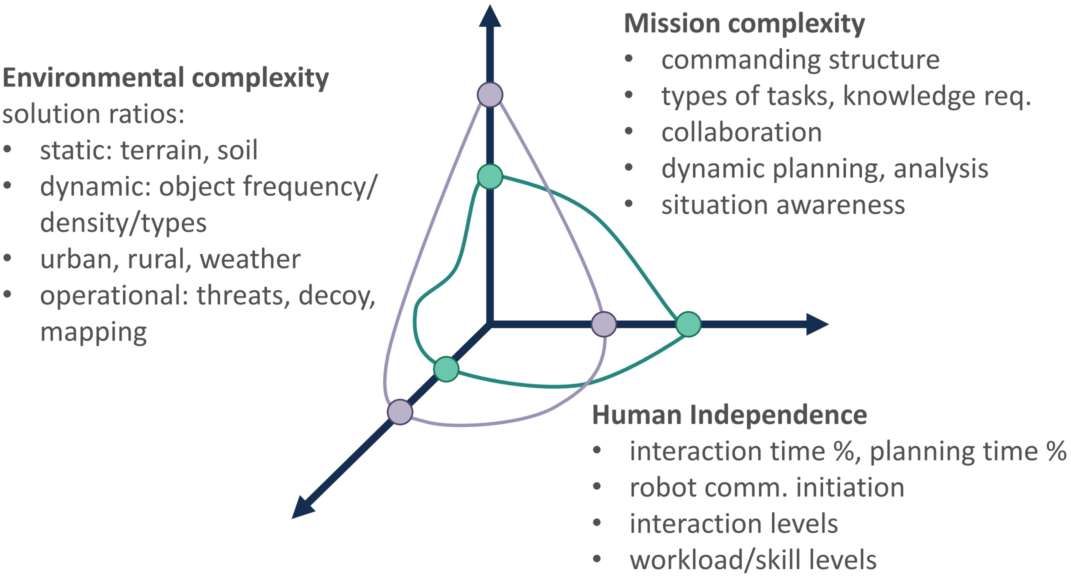


Figure 24: Three dimensions of complexities (Source: along the lines of [104])

a protective mechanism, based on risk metrics or more complex functions, to assess risks of behavioural options in the current situation. The development of risk metrics and functions to determine and control risks at run-time is primarily a research topic, but is already finding its way into standardization, as in the application rule for autonomous cognitive systems VDE-AR-E 2842-61-2:2021 [105] or ISO 21815 [106], [107], [108] on the collision avoidance of earth-moving machines.

Presentation of direct and indirect relationship between AI and safety

A few cases have already been mentioned of how AI can relate to safety, such as AI in nominal behaviour or AI for risk reduction or AI as a tool in performing the risk management process. A key differentiator between these and other cases is that AI is more or less directly related to safety A rough distinction can be made between a direct and an indirect safety relationship.

A system has a direct safety relationship when a failure or error in the system directly results in a hazardous condition for humans and the environment. This also applies to a software system that is classified as AI. Dedicated safety systems for the implementation of safety functions for risk reduction typically have a direct safety relationship. A direct safety relationship does not mean that an accident will necessarily occur if the system fails, because a hazardous condition does not necessarily cause an accident situation. As a rule, there are still the normal operating functions and the requirement situation for the safety function must also occur.

A system in which a fault or failure only indirectly leads to a hazardous condition for humans and the environment has an indirect safety relationship. Indirect safety relationships can be very complex and analyses of the causal relationships can be very time-consuming, especially in the case of interdisciplinary processes. Analysis of indirect causal chains often reveals that the probability of a hazardous event occurring is very low. However, depending on the complexity of the analysis, it is difficult to guarantee that the analysis results are reliable. In practice, therefore, direct safety relationships through systematic aspects laid out in the architecture are more decisive for the (analysis/observation of) system safety than such indirect connections.

AI WITH DIRECT RELATIONSHIP TO SAFETY

In the case of AI directly related to safety, it is essential to examine very closely whether either AI as a functional component possibly increases the level of risk or whether AI as a safeguard component really achieves a necessary risk reduction. In the following, we first present aspects that should be taken into account when considering whether the use of AI as a functional component potentially increases the level of risk. We then go on to discuss the use of AI for the necessary risk reduction and the associated quality requirements. Finally, we explain how structured safety assurance cases can be used to make the achieved quality transparent and to reveal weaknesses in arguments regarding the use of AI in the safety context.

AI potentially increases the level of risk by:

a) insufficient understanding of the system

AI offers possible solutions for cases where it is difficult to explicitly specify and program in what outputs should be generated from inputs. Machine learning allows developers to learn concepts from data and indirectly map them into an algorithm. However, the lack of explicit specification and programming of the relationship between input and output data, and the often huge parameter space of connectionist AI systems in particular, limits system understanding. Since safety demonstrations are based on system understanding, the lack of system understanding limits their use in the safety context.

b) changes to boundary conditions

In view of the fact that AI software has already been used to some extent and is expected to be used to a greater extent in operational functions and nominal behaviour in order to implement various use cases, it is quite conceivable that previous risk considerations will have to be reconsidered.

Influences on the initial risk should be investigated, but should be less critical if directly related to safety and include worst-case safety considerations.

More strongly, however, AI software could influence existing safeguards, which are dimensioned on assumptions of hazard situation probabilities, by changing these boundary conditions. It is conceivable, for example, that AI software is used to control machine cooling, but due to inadequacies, the temperatures for safety-relevant components become inadmissibly high and thus systematically run out of specification, leading to failures of these components that cannot be assessed. Scenarios are also conceivable which change the demand rates on safety systems (i.e. how often the safety system must intervene) and thus the boundary conditions of their design.

At this point it should be noted that normal software can also lead to changed boundary conditions. However, the risk or uncertainty with respect to changing boundary conditions as a consequence of the use of AI systems is usually higher: due to possibly incomplete specifications, lack of transparency of the algorithms and, depending on the concrete model, a high sensitivity with respect to parameter changes in combination with adjustments of the parameters in the context of updates and continual learning approaches (“online learning”).

c) direct influence on safety functions

AI software that is used in connection with safety functions, including its use as a protective measure itself, is occupying experts and continues to require great attention. In this context, reference should also be made to the specific use cases from industry described in this paper.

d) changes in human behaviour associated with automation

The way people interact with a technical system is changing due to increasing automation.

In the case of technical assistance systems that provide extensive support to people, the question of changes in human behaviour arises. Traditionally, this aspect is also taken into account in the risk assessment for safety. In addition to the loss of physical abilities, it may also be the case that contexts can no longer be assessed cognitively (unconscious incompetence). There is also the effect that the assistance systems might give the impression of being very reliable, which can cause a change in behaviour, e.g., due to a loss of attention on the part of the person.

At higher levels of automation, the person often still acts as a supervisor, with the system then in turn monitoring whether the person is still fulfilling their function as supervisor. AI offers special opportunities in this regard.

QUANTIFIABLE RISK REDUCTION THROUGH AI

This is the often desired and intended use of AI software, but research and development (R&D) on the reliability of AI technologies is still needed here, and challenges remain to this point, depending on the type of AI technology.

The differences between the protective measures have already been discussed, and in the field of functional safety it is common for the value of the measure to be quantified. But here, too, safety levels (SIL according to DIN EN 61508-1:2011 [101] or PL according to DIN EN ISO 13849-1:2016 [109]) are composed of quantifiable (statistical errors) and non-quantifiable (systematic) aspects. In other areas, e.g. electrical safety, measures are defined in standards but are not backed up by probabilities. Therefore, there is a debate on how to deal with AI software, and this debate needs to continue. Regardless of the outcome of the debate, it remains that the achievement of certain qualities in AI applications would be a basic prerequisite for their use a a significant protective measure.

Due to the lack of interpretability and system understanding (see a) insufficient understanding of the system) of many of today's AI systems, there is often a lack of proof procedures that guarantee (or at least make plausible) that the AI performs its task with the required reliability. Today, there is no procedure for object detection (e.g. pedestrian detection) that can reliably determine whether a pedestrian will be detected in every case. Detection methods do exist, but the statements on risk reduction obtained with them do not meet the requirements needed, for example, for perception in highly automated driving. Similarly, there is research to improve situational uncertainty assessment and handling, as well as the explainability of AI that would allow such proofs. However, even these have not yet been sufficient to provide the required proof.

There are certainly applications and technologies that are applicable based on the current state of the art, but these areas are often not seen as "true AI" by AI experts (e.g. decision trees), but still fall under the rather broad definition of AI according to ISO/IEC 22989:2022 [16] and the EU AI Act. ISO/IEC TR 5469 [33] has therefore begun to consider how AI safety applications should be classified in order to be able to derive suitable requirements and measures on this basis. This work should receive further support to arrive at a comprehensive assessment approach to AI software safety, as approaches are still incomplete, especially for "real AI technologies."

Assurance cases are a suitable framework for such a comprehensive concept, as already considered in VDE-AR-E 2842-61-2:2021 [105] and ISO PAS 8800 [110], which is currently under development.

QUALITATIVE RISK REDUCTION THROUGH AI (ASSURANCE CASES)

There is extensive experience and a consensus for the use of "normal" software in the context of safety, so that measures can be recommended with the aid of safety levels (e.g. SIL according to DIN EN 61508 [101], [102], [103], [433], ASIL according to the ISO 2662 series [455], PL according to DIN EN ISO 13849-1:2021 [111], AgPLr according to DIN EN ISO 25119-1:2021 [112]). The level is typically determined using a risk graph in which risk parameters such as the extent of damage are determined, and each parameter combination is assigned to a safety level. The safety level then determines which measures such as test methods, code review, etc. should be performed. The basis for this determination is experience. Accordingly, safety levels should only be used where there is sufficient experience. ISO/IEC/IEEE 15026-3:2022 [113] states: "Integrity levels shall be defined for an area only if a substantial body of relevant experience exists for the area that is well understood by those performing the definition." Since this experience is still lacking when using AI in safety-critical contexts, the concept of integrity levels cannot yet be meaningfully applied to AI.

However, there may be documents that compile a helpful catalogue of measures. Such a catalogue can be used to select measures. However, the question then arises as to whether these measures are sufficient for the intended application. Assurance cases are a suitable means of finding an answer to this question and evaluating whether a valid case for safety can be made. In addition to research on AI assurance measures and their collection, research on assurance cases for AI is essential to optimally exploit the current potential of AI in the safety-critical context as well, and to avoid unacceptable risks.

Assurance cases are used in system and software engineering to assure safety despite increasing the complexity of the task (e.g. along the three dimensions mission complexity, environment complexity and autonomy) and/or the use of new technologies. AI is a new technology in terms of its use in safety-critical contexts. There is little experience in this context of use and no standardized consensus on what is sufficient and when. Accordingly, such cases should also be used to guarantee safety in the use of AI.

Assurance cases consist of a structured argument, based on evidence, to assure a claim. Due to the structuring, each individual argument can be checked for validity in a dedicated manner. The use of assurance cases has been proven in other technologies, and there is broad consensus that it is also helpful in AI to prevent unsafe systems from being brought to market. Further, they favour building the necessary experience, as field data can be used to strengthen arguments and identify invalid arguments early on.

However, it should also be mentioned that while conceptually assurance cases can be established for systems with AI components, for many components it is currently difficult to prove (or at least make plausible) the evidence required for the specific assurance case. Current research approaches to structuring and testing assurance cases should be further developed and evaluated with respect to intersubjective assessability and other criteria for safety standardization.

ADDITIONAL RISK REDUCTION THROUGH AI (BEYOND THE POINT OF ACCEPTED RISK)
In this context, risk minimization can also exceed the point of accepted risk. The state of the art is to be applied for minimizing the risk and thus applications can also be used which, considered by themselves, cannot prove a defined quality of protection, i.e. those which cannot be quantified.

An example of such an application would be a measure to control compliance with correct operation ("intended use") in order to detect and/or prevent possible misuse ("foreseeable misuse") where this is not constructively possible.

Such applications, however, require close consideration of how they may work over the long term. The following case distinctions regarding behavioural changes can be usefully made here:

→ Risk reduction has a behaviour-changing effect:
With assistance systems that work well in daily use, people will tend to rely on their effects. This can be observed very well empirically in vehicle assistance systems. This change in behaviour leads to the technical system acquiring a higher value than was originally intended. This creates an effect on the original risk assessment and the effectiveness of the protective measures

→ Risk reduction does not change behaviour:
In the case of functions that do not find their way into daily use, such as AI-assisted emergency stop applications, the adaptation of humans to the system is not given because the inconvenience for such safety triggers is too high for normal operation. However, false triggering of safety systems can in turn create new hazards.

This topic is also only indirectly related to AI software. AI software is more of a technological tool to build assistance systems and autonomous systems. Behaviour change, however, is not dependent on what resources were used to build a system, but on what was built. The human behavioural changes would also occur if the same system behaviour had been implemented using a different type of software other than AI software.

AI with indirect relationship to safety
Some applications of AI with an indirect relationship to safety are discussed below. First, the application of AI in risk management will be discussed, and then AI in occupational safety.

AI IN RISK MANAGEMENT/RISK ASSESSMENT
A conceivable application of AI in risk management differs from a classic tool support as is used today for risk documentation. Smart features that perform pre-assessments of risk, where appropriate, may in turn lead the experts involved to rely more on the suggestions or the assessment than would be appropriate.

Strictly speaking, however, such systems do not fall under a safety-related tool class (and therefore have an indirect safety relationship), but aspects of behavioural change and issues of human understanding of the assessed risk of complex systems and interrelationships, as addressed earlier (see 4.2.1.2), do take effect.

It would be worth investigating whether such risk analysis tools or systems might be able to have sufficient trust and reliability potential to meaningfully complement expert assessments in a risk session, and how they should be implemented and used to avoid or address risky changes in behaviour in use.

SYSTEM COMPLEXITY, USE OF AI FOR RISK ASSESSMENTS
The increase in system complexity presents a major challenge in analyzing the critical issues in risk assessments. A wide variety of issues intersect in the possible use of AI to conduct risk assessments of such complex systems.

COMPLETENESS OF THE RECORDING OF THE NECESSARY RISK CRITERIA
Deeper analyses of risks from AI software require the comprehensive and far-reaching definition of descriptions (semantics) about interrelationships, which would have to be agreed upon across a wide variety of domains in an interdisciplinary manner.

The following complex issues are just some of the challenges:

a) Can the real world be sufficiently transformed into a digital description or is this transformation associated with an insufficiently complex representation?
b) Is there agreement on the scientific and technical relationships?
c) Are there too many different interests to reconcile?
d) How can the correctness of the criteria and also that of the derivations made be evaluated?

KEEPING HUMANS IN CHARGE AND IN CONTROL (NO SAFETY WITHOUT THE ULTIMATE HUMAN DECISION)
All safety aspects, such as completeness, correctness and liability, are tied to humans. If safety aspects are broken down to technical systems, there are strict requirements for their effectiveness and correctness. In accidents, it can be observed that human error or, technically speaking, systematic errors are often the ultimate cause. Strictly speaking, this applies to all incidents, because a technical system is developed by humans. The chain of unfortunate events or else an unexpected transgression of previously defined boundary conditions are observed when accidents are investigated.

The temptation to conclude from this that as soon as "responsibility" is transferred to the machine, the risks would decrease, ignores the disproportionately higher probability that the development of the machine was carried out beforehand with already potentially inadequate parameters and specifications on the part of a few experts.

This area should be examined in an unbiased manner. There is also the ethical question of the extent to which a machine should "independently" assess dangerous situations for humans at all. Today's safety functions such as collision avoidance in automated guided vehicles detect hazardous situations and react to avoid collisions. The system behaviour is in a certain sense "independent", because there are no user defaults for the behaviour. However, the system behaviour is comprehensible to developers and is not an end-to-end learned behaviour that gives the impression of self-determined "independent" action. Even if this type of end-to-end learning were to eventually take hold in the safety-critical context, humans would still be responsible for the consequences. In the case of explicitly programmed behaviour for dealing with hazardous situations, the ethical question arises as to which rules are permissible and which moral values apply. For example, to what extent is it acceptable to detect collision risks, algorithmically evaluate them, and make trade-off decisions? This question belongs to the new research field "dynamic risk management", which deals with enabling autonomous systems to detect and manage risks.

DYNAMICS OF CHANGES IN COMPLEX SYSTEMS
The constant change in systems will increase due to new technologies and there is the challenge for the safety domain to be able to perform the evaluation of new configurations and functionalities in a timely manner. The problems described above are again complicated by the dimension of dynamic change. The current state of affairs is that the risk analyses of the individual machine and installation components of various manufacturers are not generally visible and easily available. The reason for this is not least the knowledge (intellectual property) contained in this information. The relevant derivations from the risk analysis, which were made for the intended safe operation, are passed on in the safety-relevant accompanying documents, which are of great legal importance. Although there are key words for safety instructions in the accompanying documents, in a concrete case of damage it is the entire document, and even material from the field of marketing, which is used for legal proceedings. Only the defined exchange of this publicly made information from the risk assessment between the parties involved is not yet standardized today.

In the event of changes to the intended operating specifications set out therein, a renewed risk assessment is often required, at least in part. The frequently advertised dynamically adapting systems are state-of-the-art systems whose safety has been comprehensively evaluated by experts in all its forms and options.

AI IN OCCUPATIONAL SAFETY
All aspects of occupational safety can potentially be affected by AI and require a risk assessment (safety risk analysis is not identical to this). Occupational safety encompasses a wide range of tasks, from, for example, machine safety to chemical hazards, workplace ergonomics issues to questions of a safe route to work. In principle, it is conceivable that all occupational safety tasks could be supported by AI software tools, and the considerations are similar to those made earlier about using AI in risk management/assessment.

Conclusion
The desire and demand to use AI software, especially in safety applications, must be met rationally. An analysis that is as unconditional as possible is necessary for the respective application, and risk assessment is a central point of the open-ended consideration. The purpose of a risk assessment is not to justify the use of AI in every case, but to decide on the most reasonable safety solution. In particular, the use of certain AI software technologies in directly safety-critical applications still requires some research and development work. These results play an important role, because for high-risk applications (according to the EU proposal for AI regulation), which are also not originally safety applications, evaluation criteria from the field of safety are necessary and traceability of correct operation is required. Thus, the work

on a safety rationale for AI software can take on broad significance for other application areas as well.

It is recommended that the first steps towards safety AI software should initially be taken using simple use cases, in order to simplify the task and to be able to better evaluate the methods. Unfortunately, these simple tasks usually do not receive the necessary attention in R&D.

4.2.1.3 Standardization needs for safety

Need 02-01: Suitable definitions and regulatory criteria as a basis
Refine the AI definition in regulations in terms of safety action needs

High-risk AI systems (as defined by the EU AI Act [4]) can also be systems that are not considered safety systems. However, similar requirements do apply. Is it the legislators' wish to design all high-risk AI systems as safety systems (in the sense of fail-safe, functional safety) in the future?

Any regulation whose core aspect to be regulated is not defined cannot be applied.

Current regulation and standardization regarding safety takes software into account. From a safety perspective, the definition of an AI system should only address types of software that are not yet sufficiently addressed by current regulation and standardization.

Need 02-02: Evaluate research on safety concepts and standards
R&D on the reliability of AI technologies is necessary. Methods and procedures are needed to use this technology with confidence. The current state is not sufficient to implement resilient risk-reducing measures with AI. The first steps towards safety AI software should initially be taken using simple use cases in order to simplify the task and to be able to better evaluate the methods. These simple tasks usually do not receive the necessary attention today. There is interest here on the part of research and industry to implement these use cases, but there is a lack of funding.

Need 02-03: Promote research on safety assurance cases and evaluate standards
In addition to research on AI assurance measures and their collection, research on assurance cases for AI is essential to optimally exploit the current potential of AI in the safety-critical context as well, and to avoid unacceptable risks. Today, methods are not sufficient to provide the required evidence. The viability of safety concepts for AI must be able to be proven argumentatively based on facts (assurance case approach). In particular, the use of certain AI software technologies in directly safety-critical applications still requires some research and development work. Today, methods are not sufficient to provide the required evidence.

Need 02-04: Safety of autonomous systems
In the case of explicitly programmed behaviour for dealing with hazardous situations, the ethical (legal) question arises as to which rules are permissible and which moral values apply. For example, to what extent is it acceptable to detect collision risks, algorithmically evaluate them, and make trade-off decisions? Ethical issues are a cultural and social question and cannot be unified within the framework of standardization. Dealing with dilemma situations is less relevant in practice. Relevant are algorithmic decisions regarding the handling of risks and uncertainties (e.g., perceptions).

4.2.2 Security

4.2.2.1 Status quo

In principle, all the usual security protection targets such as confidentiality, availability, integrity or even resilience can be compromised in AI systems, as wellas in other IT systems. One speaks then of risks of the violation of a protection target for an object of protection. With DIN EN ISO/IEC 27701:2021 [128] and other subordinate standards, a standardized information security management system is already available for information security, including various measures and a risk assessment, as well as a testing and certification option. Similar standards exist in other areas or industries, for example TISAX for the automotive industry.

One example of additional risks associated with artificial intelligence is the risk of unauthorized, undetected, and targeted manipulation of training data. In a so-called "data poisoning" attack, the training data is manipulated with the aim of influencing the entire AI system by introducing an influence and thus incorrect AI model because it was trained based on the manipulated training data. Standards, e.g., of the DIN EN ISO/IEC 27000 series [131], are a basis for information security and need to be examined to see if they need to be supplemented to account also for systems with artificial intelligence. Vulnerabilities in the AI system that impact safety, security, and privacy can be much harder to identify and to remedy without transparency, traceability, and explainability. Without further information about the inner workings of the AI system, a vulnerability analysis is comparable to that of a software system using closed-box testing (ISO/IEC/IEEE 29119-1:2022 [464]), often better known as black-box testing; the vulnerability analysis is then comparatively more difficult than in a glass box test, since here the tester is only given a specification or only access to external inputs and the responses of the AI system, but no internals such as the AI model used. There is already a lot of preparatory work in terms of IT security and software security. Nevertheless, the required bridging between existing IT security standards and AI remains to be done and is a very significant challenge.

Just like IT security, AI security must also be regarded accordingly in terms of time (= over the entire life cycle) and scope (= for all components of the AI system). To stay with the example, a necessary protection against manipulation of the AI system includes protection against manipulation of the trained model and protection against manipulation of the training data, and this protection must take place throughout the life cycle: Beginning with the creation of the data and models, and continuing through their use and also during operation, AI systems´protection and its compliance must be accompanied by the appropriate monitoring of the AI system.

Status of the recommendation for action on IT security from the first edition of the Standardization Roadmap Artificial Intelligence

The recommendations for action from the first edition of the Roadmap are included again in Table 5 in order to briefly outline where there are already considerations or work in standardization and where there is still a need for action.

Table 5: Overview of recommendations for action of the 1st ed. of the Roadmap and their status quo

	Need from 1st ed. of Roadmap	Description of current status	In standardization work
1	Research/examination/evaluation of existing standards, conformity and certification procedures and existing laws	A complete research and overview has not been created yet. Standardization bodies tend not to produce overviews, but rather develop concrete criteria that are not yet available. A budget, e.g. from politics, is required for a study. The standardization and regulatory landscape on AI is also still very much in flux. An overview is given by the EU’s Observatory for ICT Standardization [115] with its Report of TWG AI: Landscape of AI Standards [116].	Yes, partially
2	Recommendations for actors and market participants	These have not yet been taken on by a standards body.	No
3	Development of supplements/adjustments in risk management	ISO/IEC 23894:2022 [25] on AI risk management developed by ISO/IEC SC 42 AI will soon be published. From a security/privacy perspective and with regard to safety issues, this standard would have to be reviewed again.	Yes ISO SC 42/CEN JTC21
4	Combining criticality levels and IT security	There is a standardization proposal from Germany at CEN/CENELEC for the classification of AI. It is planned to address risk and criticality from an AI perspective. The result is still open.	Yes CEN JTC21

	Need from 1st ed. of Roadmap	Description of current status	In standardization work
5	Define IT security criteria for training methods	This point is being dealt with together with recommendation for action 7.	Yes, partially
6	Create explainable AI	ISO/SC 42 is working on a Technical Specification on this topic: ISO/IEC TS 6254 [36] Information technology – Artificial intelligence – Objectives and approaches for explainability of ML models and AI systems. At DIN, a DIN SPEC 92001-3 [117] on the topic of explainability has currently been launched. Need for research: Still open is the initiation of basic research, which is additionally required as methods are not yet fully and widely researched and applicable.	Yes, partially ISO SC 42; DIN SPEC 92001-3 [117]
7	Define controls for IT security for AI	So far, there are studies by the European Union Agency for Cybersecurity (ENISA) [118], [119], the German BSI [81] and the Fraunhofer-Gesellschaft together with the BSI [120]. The need has not yet been taken up by a German standards body and has therefore been included again in this 2nd edition of the Roadmap in the context of testing and certification. In addition, there is the standardization requirement in the draft AI Act on cybersecurity. At ISO/IEC level, there is an initial activity in ISO/IEC JTC 1/SC 27 to confront attacks such as data poisoning with measures, but the ISO/IEC 27090 project [121] is still in its very early stages (WD stage). Activities should be strengthened for this purpose, as lists of measures (controls) suitable for AI are relevant for certification.	Yes, partially ISO Liaison SC 27/SC42
8	AI security by design and AI security by default	Security by design and security by default is required by the Cybersecurity Act. This criterion is part of a standards series for secure software development, e.g. ISO/IEC 27034 Information technology – Security techniques – Application security [122], [123], [124], [125], [126], [127]. In a possible complementary security standard for AI, this point would have to be adopted, as well as testing and certification mechanisms	No
9	Verification of the origin and protection of data	For this requirement, the AI standard series ISO/IEC 5259 Artificial intelligence – Data quality for analytics and machine learning (ML) [39] is under development at ISO/IEC JTC 1/SC 42.	Yes ISO SC 42
10	IT security of training data	All data situations of an AI require IT security measures. For activities, see Recommendation for action 7 of the 1st ed. of the Roadmap.	Yes, partially, ISO SC 42
11	Define IT security criteria for learning systems	See Recommendation for action 7 of the 1st ed. of the Roadmap.	No

	Need from 1st ed. of Roadmap	Description of current status	In standardization work
12	Verifiable identity of AI algorithms	AI algorithms would have to be provided with a verifiable identity and their function and mode of operation would have to be recorded in documentation. If possible, results should not only be shown as the probability value of a result class as the basis for a decision, but also as a confidence interval. No standardization body is currently active on this recommendation.	No
13	IT security metrics for learning systems and adversarial machine learning (AML)	There is a need for research.	No
14	Impact of availability of resources	There is a need for research.	No

Regulation, draft AI Act and standardization requirements
At the European regulatory level, more attention has been paid to the topic of security in recent years resulting in special regulations on cybersecurity. The EU Network and Information Security (NIS) Directive, which is currently being revised, has been implemented in Germany as the IT Security Act 2.0. In addition, the European Cybersecurity Act (CSA) describes further security requirements, including security/privacy by design and default, and gives ENISA more competencies and tasks. The EU Cyber Resilience Act is still in preparation, and will further increase the requirements (see Chapter 1.4).

For AI, the Artificial Intelligence Act (AI ACT) is in preparation (see Chapter 1.4). It contains instructions for action depending on a risk assessment for the use of AI, especially for high-risk AI applications. Included among others is the requirement for cybersecurity, which is always related to other requirements such as risk and quality management, logging and monitoring, transparency and information, human oversight, accuracy and robustness. In summary, a conformity assessment is expected, which should be underpinned by standards. Their development and provision should, if possible, be carried out by the European standardization organizations.

The associated draft of requirements for standardization includes a requirement for "Cybersecurity specifications for AI systems" in Chapter 2.8 [4]: The draft calls for European Standards that provide appropriate organizational and technical solutions to ensure that AI systems are resistant to changes in their use, behaviour, performance, and provide protection against a breach of security. The organizational and technical solutions should, as far as possible, prevent cyberattacks on specific AI subcomponents such as training data (cf. data poisoning attack) and ensure the security and functioning of the underlying information and communication infrastructure. The technical solutions should always be chosen according to the relevant circumstances and risks.

Status quo

Testing and certification of privacy/data protection in the use of artificial intelligence (Level 5)
The protection of personal data in accordance with the General Data Protection Regulation (GDPR) also applies to artificial intelligence. Art. 4 no. 2 of the GDPR states: " 'Processing' means any operation or set of operations which is performed on personal data or on sets of personal data." This includes the collection, recording, organization, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction of data. When these operations take place by means of automated procedures, they are referred to as "automated processing". In addition, data security must be ensured in accordance with Art. 32, which includes a risk assessment and IT security measures.

This means that the GDPR must be taken into account in every risk assessment and associated measure. In Germany, the Data Protection Amendment and Implementation Act EU

(DSAnpUG-EU) [129] was supplemented. Depending on the application and on the basis of the consent for processing, the methods of pseudonymization and/or anonymization in particular play an important role, which unfortunately are often not sufficient in the context of AI systems, since re-identifications of individuals may be possible in some cases, e.g., from the MRI (magnetic resonance imaging) scan of the head in the case of a rare disease that was treated in a specific hospital during a specific period of time. Depending on the classification of the data (personal or not and anonymized or not), the privacy of the data and the underlying individuals must be protecteddifferently . The specific data protection requirements according to the GDPR as well as the requirements from DIN EN ISO/IEC 17065:2013 [17] for conformity must be taken into account. Furthermore, standardization provides a series of standards as the Privacy Framework with DIN EN ISO/IEC 29100:2020 [133] and DIN EN ISO/IEC 29134:2020 [134] Impact Assessment and DIN EN ISO/IEC 29151:2022 [135] Guideline as well as DIN EN ISO/IEC 27701:2021 [128] (data protection along the lines of the GDPR).

An official test and certification according to the procedure in DIN EN ISO/IEC 17065:2013 [17] according to the GDPR has not yet been published, but is planned soon.

Testing and certification according to DIN EN ISO/IEC 27701:2021 [128] (when using DIN EN ISO/IEC 27001:2017 [480]) is possible, although this does not include testing and certification according to the GDPR, but supports it.

Further information is available in the Bitkom guide “Machine Learning and the transparency requirements of the GDPR [136].

For examples of existing tests and certifications in the areas of safety, security and privacy, see Annex 13.3.

4.2.2.2 Requirements, challenges and standardization needs for security

First Challenge: Definition of protection targets on the level of processes and data within the AI component
As explained, IT security targets such as integrity always refer to an object for which this protection target is to be achieved. The object in the sense of a test can be processes, data or physical components, for example AI training data. For targeted description and testing of security, it is therefore necessary and desirable to further subdivide the AI component, i.e., the system component that provides artificial intelligence. This enables to look more specifically at corresponding protection targets and thus also at individual measures for increasing protection (controls) on the basis of small and delimited areas. Staying with a concrete example: The IT security protection target of integrity should apply to the training data to provide the attacker with fewer opportunities for a data poisoning attack.

The challenge is to perform the decomposition of the AI component itself into different data or different subprocesses in such a way that existing attacks can be described, as far as possible, as a violation of protection targets of individual subcomponents. In detail, this may be different for different AI methods, but decomposing an AI component as abstractly as possible helps to describe attacks and countermeasures for entire or even multiple classes of AI methods. This then also makes it possible to prescribe the use of these measures (controls) across AI methods for the successful certification of an AI system. At the same time, a breakdown of the AI component helps to better understand the complexity and identify problems that may arise in the interaction of the AI component with the overall system. Within the life cycle of an AI system, individual subcomponents have different effects; this should also be represented by a model. Ultimately, this decomposition is also necessary from an economic point of view, as it helps to sensibly limit the scope of the corresponding measures and thus to use resources for a targeted and purposeful enforcement of security, safety and privacy protection targets.

For decomposition into AI subcomponents, ISO/IEC 22989:2022 [16] already makes a good start. This decomposition, for which there are already further approaches in research (e.g. ISTQB syllabus for Certified Tester AI Testing [137][79] (Fig. 1, page 30)), should also be pursued further in standardization and taken up and expanded for individual AI methods or method classes. The following is a proposal for discussions on an abstract component diagram that is generic enough across AI process boundaries to serve as the basis for more refined component diagrams. However, the extent to which this needs to be adapted for different AI methods should be investigated. For individual AI methods (Chapter 4.1.1.1), some AI subcomponents or process steps may be omitted from the generic component diagram. In Figure 25, the AI component is shown decomposed into different processes

[79] https://www.istqb.org/certifications/artificial-inteligence-tester

and data, which as subcomponents influence the AI functionality and are thus protected with measures to achieve the relevant protection targets in each case. Not all subcomponents are found in all AI methods. The terms in Figure 25 supplement the subcomponents described in ISO/IEC 22989:2022 [16] and are also colour-coded to the life cycle phases proposed there. The terms are further defined in Table 6.

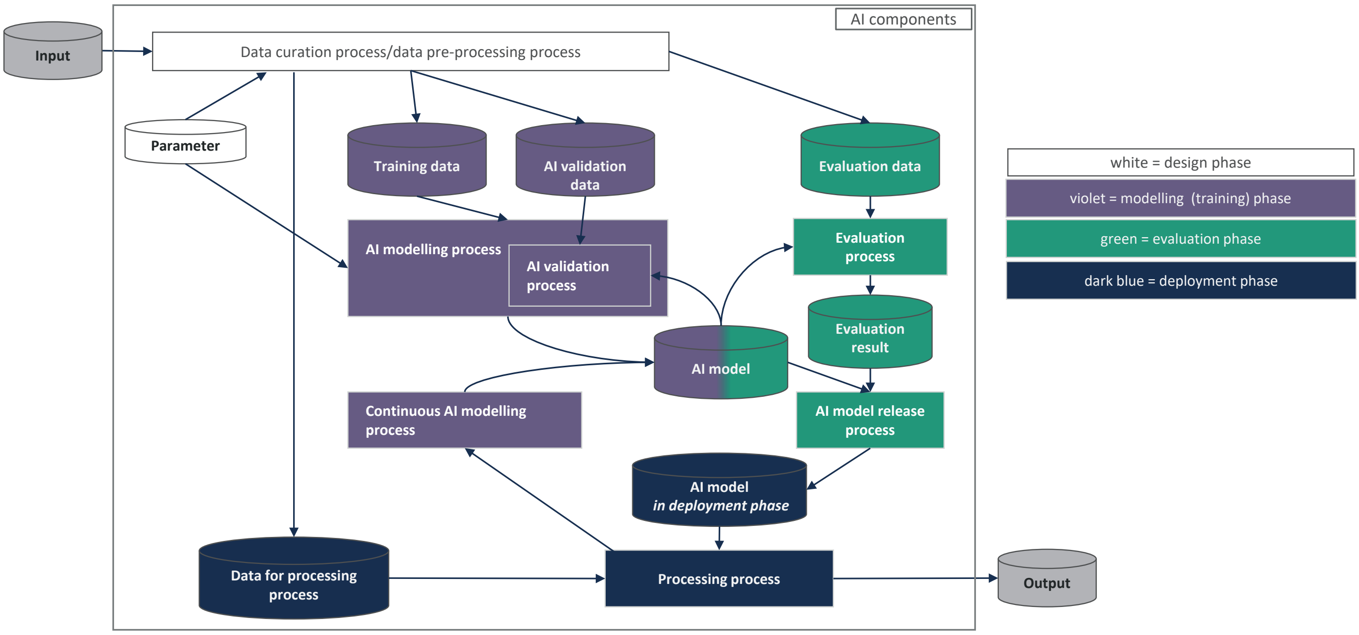


Figure 25: Component diagram (Source: Dr. Henrich Pöhls)

Table 6: Description of AI subcomponents (processes, data) that can be identified abstractly in an AI component. Not all subcomponents are found in all AI methods.

Superordinate term	Description
AI component	System component that provides artificial intelligence; consisting of several subcomponents.
AI algorithm	Totality of all processes and all data and parameters that make up an AI process
AI subcomponents (data, processes, model)	**Description**
AI model (in several phases)	Data providing the knowledge created by means of the AI algorithm AI model creation process with the help of other information such as the training data. The AI model is created during the training phase, evaluated during the evaluation phase, and only the AI model in the deployment phase is used by the AI processing process to generate the output of the AI component.
Processing	Process that generates an output by means of the data for the processing process and the AI model
Training process / AI modelling process	Process that generates a new AI model or extends an existing AI model using the training data and other inputs such as an AI model evaluation result, hyperparameters, or internal model parameters. If it is a continueous AI modelling process, then information from the processing process (results, internal values, but also data for the processing) also flows in, for example as part of an AI process that continues to learn (continual learning).
AI validation process	Process that uses validation data and other parameters to evaluate an existing AI model and drive the AI modelling process
AI evaluation process (within the context of the test process)	Process that evaluates an existing AI model using the evaluation data and produces a model evaluation result
Data curation process/ data pre-processing process	Process for transforming the input (raw data) into a representation suitable for the application of the processes (AI training process, AI processing process, AI validation process) of the AI algorithm
Training data	Data for the training process, generated by a data curation/data pre-processing process
AI validation data	Data for the internal evaluation of the training process generated by a data curation process/ data pre-processing process
Test data	Data for the AI evaluation process, generated by a data curation/data pre-processing process.
Input (raw data)	Unprocessed data passed to the AI component as input from the AI system, for example, image content in the case of an image recognition AI component
Data for processing process	Data prepared by the data curation/data pre-processing process that is fed to the AI processing process to produce an output
Output	Results of the AI components, which were determined by applying an AI algorithm and an AI model from the input

Table 7 below maps processes and data of abstract subcomponents to life cycle phases.

Table 7: Life cycle stages along the lines of ISO/IEC 22989:2022 [16]

Life cycle	Description
Design phase (see Figure 25 white marking)	Development of the parameters and selection of the AI process; the processes required to generate the parameters fall into the inception stage (ISO/IEC 22989:2022 [16]), and partially into the design and development stage.
Modelling (training) phase (see Figure 25 violet marking)	Creation of an AI model; partly includes the design and development phase according to ISO/IEC 22989:2022 [16].
Evaluation phase (see Figure 25 green marking)	Includes, but is not limited to, the evaluation process and the AI model release process; these processes are to be located in the verification and validation, continuous validation, re-evaluate phases of ISO/IEC 22989:2022 [16], as appropriate. (Note: ISO/IEC 22989:2022 [16] uses the term "validation" here – this has several meanings in different but here relevant contexts, see Chapter 9 and Chapter 4.4.2.3 for more details).
Deployment phase (see Figure 25 dark blue marking)	Use of the AI model to generate outputs based on inputs; corresponds to the deployment phase according to ISO/IEC 22989:2022 [16].

ATTACKS ON THE AI SUBCOMPONENT AS A VIOLATION OF PROTECTION TARGETS

AI attacks should be understood as violations of protection targets for AI subcomponents whenever possible. New special attack possibilities on the data such as a data poisoning attack or attacks to extract training data from trained models (privacy attacks) can then also be reduced by IT security measures.

In this regard, many attacks from AI systems depend heavily on the underlying data. Especially in the case of personal data, the attacks target privacy accordingly. In the case of non-personal data, the primary concern is the economic damage caused, for example, by the manipulation of systems or by the disclosure of secret data that was deemed to be used only for training, for example, but that can later stil be obtained from the trained model by the adversary. Therefore, the various data and the processes associated with them are a primary object of protection. Risk analysis and data management should, of course, always take place throughout the entire life cycle.

The first step is to analyze existing standards on security management (DIN EN ISO/IEC 27000 series [131]), life cycle, function representation, modularization, secure software design (ISO/IEC/IEEE 29119-2:2021 [465], DIN EN ISO/IEC 27037:2016 [130]) and AI ecosystem in terms of security and privacy in AI; in addition, the analysis of AI standards in progress by the bodies CEN CENELEC JTC 21 and ISO/IEC SC 42, ISO/IEC 27090 [121], ISO/IEC TR 27563 [138]; furthermore, the analysis of current regulations related to AI such as the EU Cyber Security Act (CSA), the ENISA study "Securing Machine Learning Algorithms" [119], the Network Information Security Directive (NIS Directive) and the planned Artificial Intelligence Act (AI Act) including the standardization requirements. The studies of the German BSI [81] and the Fraunhofer-Gesellschaft [120] with examined criteria and processes also support the development of a standard with manageable testing and certification criteria and processes.

Need 02-05: Abstract decomposition of the AI component into data and processes

Further refine current components of an AI system, building on the current state as in ISO/IEC 22989:2022 [16], (in line with current research and the proposal) and decompose the components to accurately describe attacks and vulnerabilities. The aim is to provide an abstract component model for further use in describing risks and measures for various AI processes and for AI certification.

Need 02-06: Match existing AI attacks and risks with existing certifiable IT security objectives
If one creates a mapping of attacks on AI components (e.g., data poisoning) to IT security protection targets according to a description of the AI components' objects worthy of protection, this enables existing building blocks from the testing and certification of IT systems to also be reused for AI systems as quickly as possible. As a basis for such a mapping, the existing documents of ENISA [119] or BSI (reference cloud AI catalogue) [81] (the latter with ISO SC 38, if necessary) should be pursued further and fed into standardization in a manner that is as free of contradictions as possible between ISO/IEC SC 27 (IT security) and ISO/IEC SC 42 (AI). There are already testing processes and corresponding certifications for IT security. Where possible, these should also find application for testing and certifying the IT security of the AI system or the individual AI components in use for the entire system. In order to avoid unnecessary descriptions of new processes and controls for AI systems and the AI component(s) used there, it is necessary to describe existing threats to AI components in terms of the object of protection (if necessary, also only for subcomponents of the AI component such as data, model, process, etc.) and the IT security protection target (for example, integrity). This would then allow certain controls to be reused, for example data governance leads to an overview of where data comes from, thus making attacks on the integrity of training data more difficult and reducing the risk of a "data poisoning" attack. This would enable an initial catalogue of measures (as in the DIN EN ISO/IEC 27001 [480] Annex or in DIN EN ISO/IEC 27002 [481]) to be drawn up for AI security and AI privacy, based on existing measures. This also reveals potential gaps, i.e., protection needs for which AI-specific measures are required. Where the attack vectors are very specific and cannot (or cannot easily) be mapped to a set of existing IT security protection targets, specific criteria must then be developed.

Need 02-07: Standardization of AI product and process testing procedures for security and privacy
IT security and privacy for AI are both issues of an AI security management system in the organization, across the life cycle and supply chain, as well as from a functional product perspective of a singular software component or from the perspective of the comprehensive AI system complex including the possible interactions. Security and privacy standardization with appropriate control criteria, test tools and test methods, as well as management system requirements for testing and certification should be developed for all areas, especially for machine learning methods and in critical environments/infrastructures. There are various established test methods and certification schemes for testing the IT security of products, systems and processes. New approaches are under development to adapt to the changing challenges in IT security. Test methods and accreditation methods are essential to ensure the quality of testing by independent third parties and to improve the traceability and comparability of results.

As suggested in the ENISA report [119], further research should investigate and validate adapted security controls for machine learning, establish benchmarks for their effectiveness, and standardize them with respect to their implementation.

Second Challenge: Elaboration of a horizontal cross-sectional standard and vertical manifestations
Over the years, security standards for testing and certification and, more recently, approaches to AI standards have developed in a wide variety of sectors and fields of activity. However, from the company's point of view, it makes sense to work with as few comprehensive and recognized standards as possible. A generally applicable horizontal standardization of cybersecurity and privacy for AI would be very helpful from the point of view of industry, as would the possibility of having supplementary standards available in the event of sector peculiarities.

For example, for medical devices, the Food and Drug Administration (FDA) has proposed a Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) [139], a framework for how continual learning systems could be tested and approved. The framework calls for SaMD (software as a medical device) Pre-Specifications (SPS) to describe the foreseen changes (in terms of "performance," "inputs," or "intended use") that may be considered permissible for the system so that a new certification is not required in the event of changes/continual learning. It also requires an algorithmic change protocol (ACP) so that appropriate testing can be used to demonstrate that the risks that can be considered acceptable under the SPS are adequately controlled. In a certain sense, an automated revalidation of the system takes place via the ACP.

Need 02-08: Elaboration of a horizontal cross-sectional standard and vertical manifestations on security
It is recommended that horizontal topics on cybersecurity and privacy for AI be worked out for testing and certification

that affect all sectors, as well as an interface with sector-specific requirements. For example, one horizontal issue would be the requirement for appropriate access control. Special security requirements from the sectoral environment, such as those for medical devices, can be viewed as a vertical form.

Third Challenge: Development of metrics and controls according to the standardization requirements of the EU AI Act

The proposed AI Act includes various cybersecurity requirements. Therefore, in the draft standardization request, a standardization of cybersecurity related to AI is included. The required standards are expected to be available along with the enactment of the AI Act beginning in 2024.

Need 02-09: Development of metrics and controls according to the standardization requirements of the planned EU AI Act

Develop standardization on cybersecurity requirements from the AI Act for metrics and controls to measure and prevent cyberattacks, and methods for testing, auditing, and certification, including requirements for criteria for audit measures and auditors.

In this context, it seems important to establish a joint working group with the cybersecurity and AI bodies in the standardization organizations of Germany, the EU, and possibly also internationally.

Fourth Challenge: Test criteria for testing tools on cybersecurity and privacy for AI

Currently, there are no testing tools and testing criteria for cybersecurity and privacy testing tools for AI. Because AI systems are fundamentally more complex IT systems, there is a corresponding overlap in the application of testing tools to IT security and privacy.

Need 02-10: Test criteria for testing tools on cybersecurity and privacy for AI

As far as testing the AI component/algorithm is concerned, there is still a lack of testing tools and methodologies. Tools for testing AI-specific criteria, as well as suitable test criteria for the testing tools themselves should be developed or existing methods for testing IT security should be supplemented accordingly.

Fifth Challenge: Quantifying robustness for machine learning models

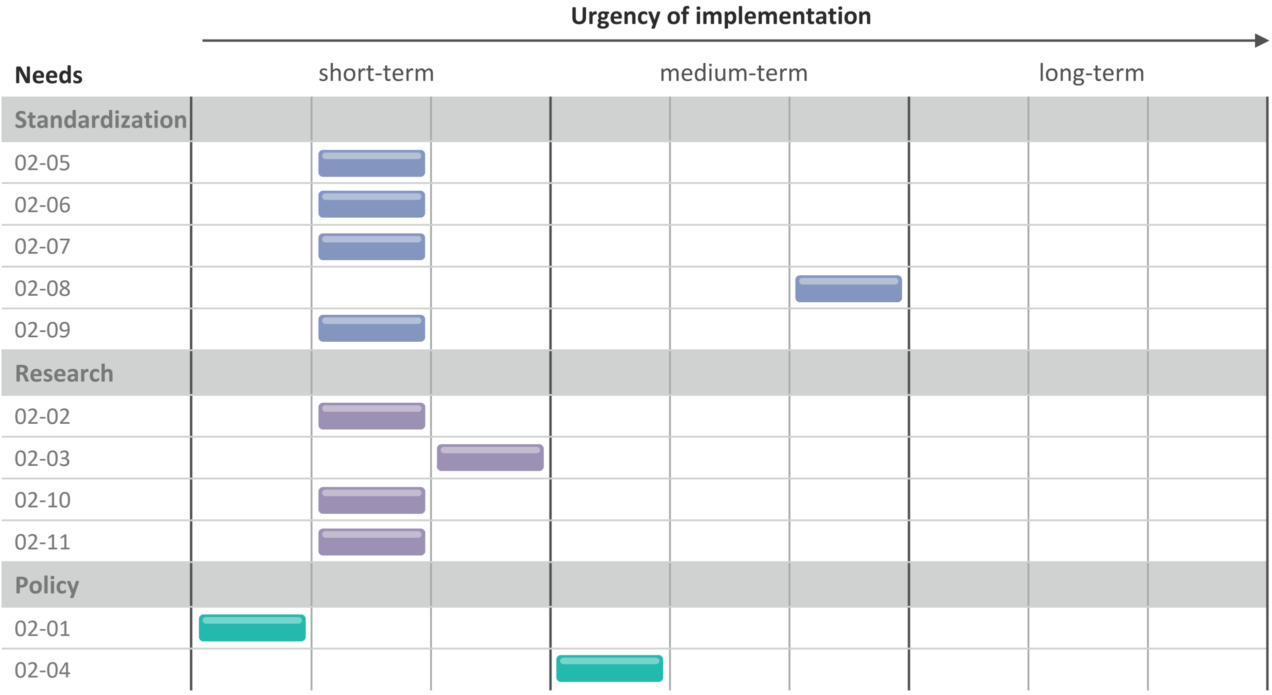
One of the objectives of the certification of AI-based systems is to quantify robustness. Two types of robustness will be considered here: (1) robustness against naturally occurring perturbations of the input data, and (2) robustness against special attacks, e.g., adversarial examples. Methods and schemes are to be developed for a corresponding robustness certification, the results of which will be incorporated into the certification process and will contribute to an appropriate assessment of the safety of the overall system.

The challenge in quantifying robustness for AI models lies in choosing suitable methods. Elaborate approaches to empirically measure the robustness of models against attacks already exist, which are based on state-of-the-art attack methods [140]. However, these approaches are not yet applicable for all models, architectures, and use cases. Also, current methods do not provide reference values for ranking robustness values.

Need 02-11: Quantifying the robustness of machine learning models

Based on the above-mentioned challenges, the corresponding recommendation for action and the following need for research arise: Further methods for quantifying the robustness of AI models should be developed. These should be included in a potential standardized certification process. The future methods should allow a measurement of robustness independently of the model architecture and other properties of the system. Thus, the applicability should remain guaranteed even for large models. New methods should also allow a relative classification of the robustness of the model as well as of the overall system, for example, by means of suitable reference values.

The Working Group Security/Safety ranked the identified needs according to the urgency of their implementation. Figure 26 shows the urgency of implementation, categorized according to the target groups of standardization, research and policy.



Urgency of implementation →									
Needs	short-term			medium-term			long-term		
Standardization									
02-05		■							
02-06		■							
02-07		■							
02-08						■			
02-09		■							
Research									
02-02		■							
02-03			■						
02-10		■							
02-11		■							
Policy									
02-01	■								
02-04				■					

Figure 26: Prioritization of needs for the key topic Security/Safety (Source: Working Group Security/Safety)

4.3 Testing and certification

Basically, the assessment of the quality of an AI application can be made from three perspectives:

1. **Societal-normative assessment:** The first perspective relates to the question of whether the use of an AI application is consistent with societal values. This perspective is addressed primarily in the chapter on sociotechnical systems (Chapter 4.4).
2. **Sectoral, deployment-specific manifestations:** This perspective refers to whether the AI application being evaluated meets the requirements of its deployment environment.
3. **Horizontal, application-agnostic technique consideration:** This perspective mainly refers to the evaluation of the AI technologies used and makes statements about the robustness of an ML model, for example.

The difference between the third and second perspectives is that an AI component (such as a specific ML model) is usually embedded in a larger (IT) system (e.g., a highly automated vehicle). The sectoral requirements (e.g., a safety or security requirement for the highly automated vehicle) relate to the overall system, while the horizontal requirements relate to an individual AI component or technology. These requirements for the overall system must then be broken down into requirements for the individual components, e.g., for the AI components or for individual datasets (see also presentation on safety arguments, for a concrete example from the field of highly automated driving, see for example [141]).

The normative requirements essentially reflect the socio-political duty of care. Conformity with societal, ethical and legal frameworks mainly serves to protect legal rights and requirements that go beyond them, which result from ethical and societal debates [68]. As a key technology of digitalization, AI is permeating many areas of life and work. AI conformity testing in these categories is intended to prevent and help avoid harm to groups and individuals, injustice, or ethically unjustified conditions of society. This requires the anchoring of fundamental principles of action all the way into AI technology. The ethical requirements for AI

- → formulate principles of the overarching framework for action,
- → are based on social and political consensus,
- → define the dimensions of the actors‘ responsibilities, and
- → should be standardized throughout Europe in an overarching political harmonization process.

Normative or ethical requirements for AI address questions about the pros and cons of using AI, such as using medical diagnostic systems as decision-making systems that replace specialists or as systems that assist specialists in their decisions (see [143] et al.). Quality labels have been proposed specifically for ethical considerations, which are based on value analysis procedures from a combination of target criteria, indicators and measurable variables (Chapter 4.8.3). All assessments of normative requirements should be able to be based on technical tests. Normative or ethical requirements for AI are considered in more detail in the chapter on sociotechnical systems (Chapter 4.4).

The application-specific requirements translate the normative principles into concrete application and add specific application requirements. They

- → form the basis of risk classification, e.g. according to the AI Act,
- → address relevant ethical aspects in the process,
- → use the New Legislative Framework to provide component-by-component support for manufacturers‘ presumption of conformity (Chapter 4.3.1), and
- → formulate requirements in principle for the entire technical system in which AI is embedded.

Application-specific requirements are at the heart of the vertical evaluation of AI, which assesses whether the AI is suitable for a specific application, for example, evaluating whether the tested accuracy of a diagnostic system is suitable as decision support for a radiologist, or whether a highly automated vehicle meets safety requirements.

Finally, the technical requirements for an AI are used to verify that the AI application is correctly specified, developed, and operated. Here, for example, certain measures are used to determine how accurately the diagnostic system classifies, i.e. how correctly it assigns X-ray images to pathologies. Technical requirements for an AI

- → formulate technically testable requirements across applications on a horizontal level,
- → focus on AI technologies and their technically motivated typical uses,
- → offer a spectrum of test methods with different test depths for selection, and
- → enable the testing of hybrid and embedded AI in concrete deployment scenarios.

This results in a three-stage cascade of requirements from the ethical to the normative to the technical level (see Figure 27), whereby the classification into risk groups is sensibly carried out at the application-oriented level, but the actual technical verification of the requirements is performed at the horizontal level.

The cascade and relationships and responsibilities must be addressed for the successful implementation of legal frameworks for action. For example, in responding to the European Commission's standardization requirements related to the proposed Artificial Intelligence Act (AI Act) in the current version, there is much leeway to propose standards at all of the above levels and with any hybrid forms. The requirements cascade is also reflected in the organization of national and international standardization bodies for AI (Chapter 3.2). Working groups and projects for vertical, i.e. application-related standards make conceptual use of horizontal, cross-application standards. And this Roadmap also follows this approach: The sectoral studies build on the basic technological and testing methodology aspects.

The consensus in standardization bodies and expert groups is as follows: There is an urgent need for clarification and action to establish test methods at the different evaluation levels and to make quality assurance for trustworthy AI in industry and society transparent. This chapter provides insight into the various dimensions of implementation, formulates the resulting issues as standardization needs, and concludes by bundling them as an urgent recommendation of developing and establishing a horizontal, cross-application AI certification programme.

4.3.1 Status quo

A programme for developing cross-application AI certification can have great appeal if it is compatible with existing conformity assessment methods and quality infrastructures. In the following, the term "AI certification" refers to a toolbox that includes different types of conformity assessment that may come into play as evaluation activities in the context of AI. The naturally brief description of the status quo here focuses on a few key questions, such as:

→ What shape and scope can AI certifications take?
→ Which quality dimensions of an AI certification can be identified? How can they be classified and, if necessary, also related to regulatory requirements?
→ How is the object of conformity assessment to be identified and selected?
→ What types of conformity assessment are relevant? Which test and inspection procedures and validations play a role?
→ What can AI certifications look like? How can they be applied to existing horizontal standards?
→ What vertical standards can be used to implement testing and inspection procedures and validations? Which ones need to be further developed and which ones need to be newly developed? (See also AI suitability in Chapter 3.3).
→ How do users, providers, manufacturers, and developers benefit from the proof of AI trustworthiness? What contribution can a cross-application certification procedure make to the acceptance of AI in industry and society?

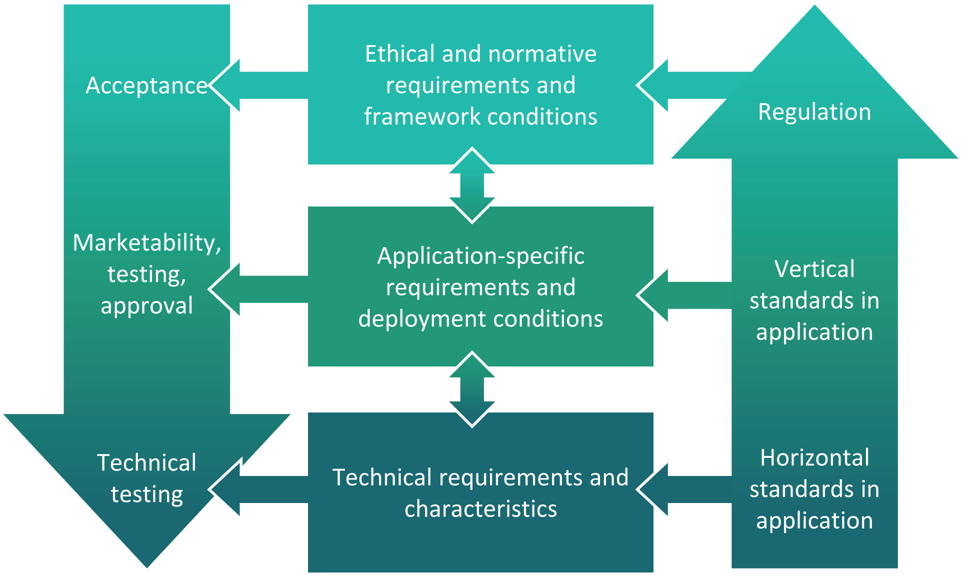

Acceptance
Marketability, testing, approval
Technical testing
Ethical and normative requirements and framework conditions
Application-specific requirements and deployment conditions
Technical requirements and characteristics
Regulation
Vertical standards in application
Horizontal standards in application

Figure 27: Three-level requirements cascade (Source: BSI)

This sub-chapter introduces important concepts and terminology that form the basis for further discussion of the topic of "testing and certification" of AI systems. For this purpose, it is discussed at the beginning (Chapter 4.3.2.1) which entities (such as systems, organizations, persons, etc.) can be the object of an AI certification. Since the discussion should build as much as possible on the established concepts of conformity assessment, the following (Chapter 4.3.2.1) introduces important principles and concepts of conformity assessment. The sub-chapter ends with a presentation of the most important quality dimensions for trustworthy AI (Chapter 4.3.2.1).

4.3.1.1 Regulatory requirements

There are a number of international and national regulations, three of which are considered as being particularly significant for conformity assessments of AI applications, services and systems for the Standardization Roadmap Artificial Intelligence:

1. the European General Data Protection Regulation (GDPR) with the implementing or accompanying standards for a data protection certificate and data protection management,
2. the European Machinery Directive (to be replaced by the Machinery Regulation in the near future) and the German Product Safety Act and their implementation with a focus on accident prevention, and
3. the European Commission's draft regulation establishing harmonized rules for artificial intelligence.

The European regulations mentioned as examples are part of the implementation of the "New Legislative Framework" (NLF). This is a package of measures to improve market surveillance (Decision of the Parliament of the European Union (EU),768/2008/EC [144]) and for the placing on the market of industrial products (Regulation (EC) No 1025/2012 [169]) in the member states, as well as to increase the quality of conformity assessment through clear rules of accreditation (Regulation (EC) No 765/2008 [145]). With the entry into force of this latter Regulation under this package of measures (the NLF), accreditation is a sovereign task throughout the EU and is performed in the respective member states by a single national accreditation body. In Germany, the Deutsche Akkreditierungsstelle (German Accreditation Body) (DAkkS) is the competent authority. For independent conformity assessment bodies of the first, second or third side with their head office in Germany, this means that only accreditation by DAkkS is permitted. Details of these regulations are explained in Chapter 1.4 and Annex 13.

Most of the EU regulations and directives relevant to AI expect risk-based testing and, if necessary, certification for defined high-risk applications and harmonized standards suitable for this purpose. The aim of this chapter is to develop recommendations for these and other requirements for testing and certification.

4.3.1.2 Ensuring the competence of organizations and protecting consumers

For example, a working group of the international standardization organizations is developing an International Standard ISO/IEC 42001 [27] for AI Management Systems (AIMS). An AIMS supports companies, organizations and institutions. This should define suitable strategies and processes for the trustworthy development and use of AI systems. The aim is to increase trust and acceptance of AI as a key digitalization technology. The development of AI and especially of automated decision-making processes leads to challenges regarding consumer trust and welfare.

Consumer protection's view of learning systems is naturally critical. Since learning algorithms can process data with a precision and speed that humans can no longer comprehend, consumer protection points to associated risks, especially when decisions are made without the results being verified by humans. A major problem is the distortion of relevant data. Machine learning is based on the recognition of patterns within datasets. Problems arise when the database does not form a representative cross-section and distorts the learning processes. This problem is addressed in particular in the test dimension "Bias, fairness and avoidance of undesirable discrimination" (cf. Chapter 4.3.2.1).

Consumer advocates also point to the potential consequences of such biases in algorithmic decision-making (ADM) systems as specific AI systems. In many cases, a decision made with the help of such systems can have a significant impact on individuals, for example in the credit industry, on the labour market, in healthcare, or in legal disputes. A European Parliament Decision calls on the European Commission to investigate whether there is legal certainty for consumers in a world increasingly influenced by AI and automated decision-making.

The German Federal Ministry of Justice, and for consumer protection, wants to advance the development of AI systems in consumer protection. The Ministry specifically promotes the development of AI applications with the programme for promoting innovation in consumer protection. The focus is on AI-based application scenarios and prototype solutions that make everyday life and self-determination easier for consumers, which are designed with the target group in mind, enhance quality of life and contribute to consumer protection.

4.3.2 Requirements and challenges

4.3.2.1 Basic concepts

Objects of conformity assessments
This chapter explains which entities can be the object of a conformity assessment. For this purpose, the relevant entities are introduced, supported by examples, and relevant standards are referenced.

In general, conformity assessments include, but are not limited to, testing, inspection, validation, and verification.

In the following, the terms "certification" and "testing" are used non-technically as synonyms for the various types of conformity assessment as well as their evaluation activities.

Because of the breadth of use and technical complexity, it is appropriate to differentiate between the various objects of conformity assessment that are to be evaluated for compliance with their requirements. The technology-specific test objects of conformity assessments (such as an audit) are the AI systems themselves in the sense of software with or without hardware components. It is easy to see that AI quality assurance considers such products, systems and solutions as the object of conformity assessment. Their actions have direct effects on the AI itself or on the environment. All other entities, e.g., persons and organizations, (information) technical systems, infrastructures, etc., may be closely linked to the AI solutions, but as a rule they bring their own impact potentials that go beyond AI. From an AI perspective, they are rather higher-level (indirect) objects of conformity assessment in the field of AI. These usually also have their own specific standards for their conformity assessment (for their "testing"), such as an AI management system for companies and organizations.[80]

TECHNOLOGY-SPECIFIC OBJECTS OF AI TESTS
The quantity and nature of the requirements mentioned at the beginning of this document naturally depend on the object of the conformity assessment. In this context, the conformity assessment of AI must take into account the increasing digitalization in industry and society. AI is increasingly finding application in various everyday products, in complex technical systems in industry, and in special information technology solutions. The main object of an AI conformity assessment is software that contains AI components, such as a module based on machine learning. The essential point in the discussion is that classic concepts for software quality assurance and testing fall short in the case of AI. Furthermore, the implementation of the software in corresponding appropriate hardware can play an important role in the assessment. Prominent examples here are cloud environments, edge computing or, as a special case, neuromorphic hardware. In general, assessing the performance of AI-based software can depend heavily on the implementation.

In view of the requirements described at the beginning, four technology-specific categories of AI test objects can be distinguished in particular.

1. **AI applications:** This term is used to describe an AI-based software solution, which can be part of a larger IT system (see also point 4.) For example, an application for credit scoring or anomaly detection can be thought of here.

Figure 28 illustrates an ML-based AI component and its application.

2. **AI services:** This refers to an AI application that is provided as a software service. A typical example is certain basic AI services (e.g., optical character recognition (OCR) systems), which are provided by large cloud service providers, for example. AI services can be implemented as cloud-based solutions, where the cloud can be private or public, or via hybrid cloud edge systems.

[80] A list of relevant vertical testing standards can be found in the following chapter.

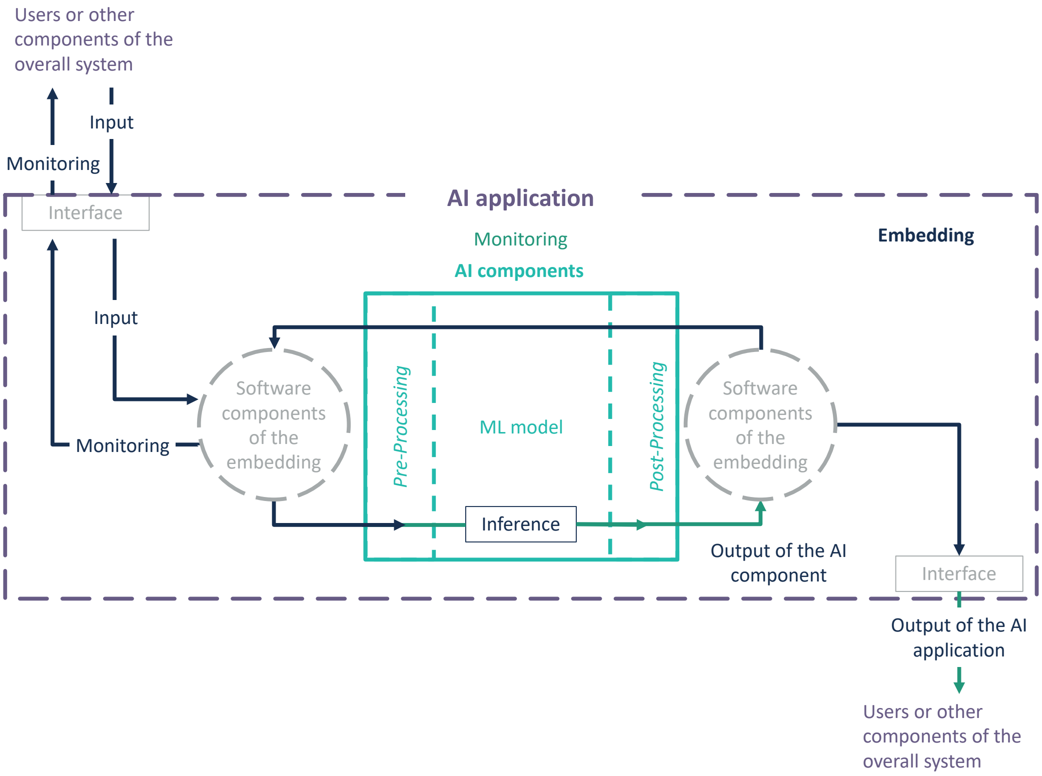

Users or other components of the overall system
Monitoring
Input
AI application
Interface
Monitoring
AI components
Embedding
Input
Software components of the embedding
Monitoring
Pre-Processing
ML model
Post-Processing
Software components of the embedding
Inference
Output of the AI component
Interface
Output of the AI application
Users or other components of the overall system

Figure 28: Representation of an ML-based AI component (Source: along the lines of [120])

3. AI module: An AI module refers to AI services as building blocks in a chain of delivery relationships involving multiple IT components or AI services. Technically, the AI module is no different from an AI service; the definition here emphasizes the importance of such components as an important part of the AI supply chain and the related issue of accountability for the quality of the overall AI system.

4. AI system: Finally, an AI system is understood to be an overall IT system that contains one or more AI applications as embedded components. It should be noted that AI systems are usually hybrid, meaning that the intelligent behaviour is realized via the interaction of several AI components and other classic software modules, whereby a variety of other AI methods can be applied to the AI components in addition to machine learning. The risks of systems with AI components must be considered when testing AI components.

Life cycle of AI systems

Due to the breadth of use of AI systems, the embedding of AI components in complex technical systems and the diversity of technologies in partly hybrid AI systems, AI standardization takes the entire life cycle of an AI system into account [16][81].

Thereby the phases

→ inception,
→ design and development,
→ verification and validation,
→ deployment,
→ operation and monitoring,
→ continuous validation
→ re-evaluate and
→ retirement

explicitly relate both to dimensions of AI trustworthiness (see below) and to AI assessment processes, e.g., the continuous risk management process.

Standards and specifications from other areas relevant to AI quality and conformity assessment exist for the development and operation of AI systems. For example, certain AI processes can be integrated into existing standards created for software development, such as ISO/IEC/IEEE 12207:2017 [148], ISO/IEC 27034 series [122], [123], [124], [125], [126], [127], as well as ISO/IEC 25010:2011 [152] and ISO/IEC 25059:2022 [35].

Data-driven systems, for example, can be characterized by a high degree of adaptation to their environment during the operational phases. During operation, the life cycle model provides for continuous monitoring and validation of the AI system as special monitoring processes. Various methods can be considered for monitoring learning systems, specifically.

[81] A detailed look at the AI life cycle using different models can be found in the “Basic topics” chapter (Chapter 4.1.2.3).

The MLOps process (machine learning operations process) deserves special mention. MLOps repeatedly deploys and continuously monitors machine learning models. This optimizes the models in productive use when the database changes. For an introduction to MLOps processes, see for example Beck et al [153]. Within the framework of conformity assessment, a continual inspection (embedded audit agent) in the sense of DIN EN ISO/IEC 17020:2012 [157] can be implemented for such processes in the future, if necessary.

The AI Act and the European Commission's standardization requests call for horizontal, cross-application conformity assessment procedures (such as AI certifications) for the technology-specific test objects – AI applications, AI services, AI components, and AI systems, and for the corresponding development processes and life cycle observation and ongoing assessment. It is therefore important, as part of the corresponding evaluation activities of conformity assessment such as inspection, testing or validation and verification, to develop appropriate testing standards and testing fundamentals at Level 4. Development and standardization must be designed in the short term, established in the medium term, and continuously adapted in the medium and long term based on technological progress and the growing range of applications.

SUPERORDINATE (INDIRECTLY) RELEVANT TEST OBJECTS
According to ISO/IEC/IEEE 12207:2017 [148], ISO/IEC 27034-1:2011 [122] and ISO/IEC 25010:2011 [152] together with ISO/IEC DIS 22989:2022 [16], derived questions for the verification of quality characteristics result from the phase view.

1. Which minimum standards for organizations, institutions and – in the case of distributed AI systems – infrastructures should be reviewed during the development and ongoing operation of AI applications?
2. Which roles of persons in the development and ongoing operation of AI applications can be identified from the above standards and how are the corresponding qualification profiles, e.g. for AI quality officers, to be examined? Can requirements for personal certification, e.g. for AI developers or AI quality auditors, be derived from the quality assurance process?
3. Quality infrastructure What requirements must testing laboratories and certification bodies involved in the conformity assessment (e.g. testing or certification) of AI systems meet?

The AI management standard (ISO/IEC 42001 Information technology – Artificial intelligence – Management system [27], [142]) addresses the first question. In order to prove that the requirements listed there have been met, a management system certification according to DIN EN ISO/IEC 17021-1:2015 [22] must be carried out by an independent certification body. As part of this certification, the management system with its defined processes and distribution of roles, as well as the organization's competence to manage and operate this system in compliance with the standards, is analyzed, evaluated and assessed in a scientifically reproducible manner. Certification of a management system is usually carried out at the request of the customer. In doing so, the certification body is to ensure that all standards required for certification are applied.

With regard to the second set of questions, reference is made by way of example to the certification of persons by the German Federal Office for Information Security (BSI) as part of the implementation of certification programs for IT security. The BSI carries out certifications of persons in accordance with Section 9 of the BSI Act, and accredited conformity assessment bodies can certify persons in accordance with DIN EN ISO/IEC 17024:2012 [155] because qualified persons are required to carry out evaluations and tests for the purpose of certifying products and management systems and to support the BSI in the area of IT security services. Likewise, qualification levels may be set in specific international standards. For example, a subordinate standard to DIN EN ISO/IEC 17021-1:2015 [22] may specify specific qualification requirements for certification bodies and auditors that audit and certify AI management systems within organizations. The goal of such a process for AI is to provide competent individuals in the areas of application and to ensure the quality and comparability of evaluations, audits, and services. The qualification requirements as well as training programmes for AI developers, auditors, quality representatives and special users are to be developed in the short term together with the testing standards at Level 4.

The third set of questions leads directly to the framework and procedures of the quality infrastructure. They are briefly explained below.

Types of AI conformity assessment
AI tests can basically be understood as conformity assessments based on one or more conformity assessment standards, which describe the scopes, need-based test criteria, requirements and proof, method and management for conducting the assessment. An important categorization in conformity assessment is how the person performing the as-

sessment relates to the object of the conformity assessment. There are three types of conformity assessment defined in DIN EN ISO/IEC 17000:2020 [147].

1. **First-party conformity assessment activity:** This conformity assessment activity is carried out by the person or by the organization that is the object of the conformity assessment or that offers it.
2. **Second-party conformity assessment activity:** This conformity assessment activity is performed by a person or an organization having an interest in the object of the conformity assessment as a user.
3. **Third-party conformity assessment activity:** This conformity assessment activity is performed by a person or an organization that is independent from the provider of the object of the conformity assessment activity and has no interest as a user.

Insofar as requirements are specified with the aim of demonstrating their fulfilment, these must also be suitable for such conformity assessment. Furthermore, according to ISO/IEC Directives, the requirements must apply to the particular object being evaluated. Details of implementation are specified separately, e.g., evaluation procedures (e.g., test methods), competency criteria, and other requirements for the conformity assessor.

The “neutrality principle” applies, according to which the requirements must be formulated or separated in such a way that it does not matter who determines and evaluates their fulfilment. These can be internal bodies (first-party conformity assessment), potential buyers/users (second-party conformity assessment) or independent entities (third-party conformity assessment). Also according to ISO/IEC Directives, no new conformity assessment bodies may be implemented by sectoral TCs. Accordingly, the requirements must be able to be evaluated or applied by a testing laboratory (according to DIN EN ISO/IEC 17025:2018 [156]), an inspection body (according to DIN EN ISO/IEC 17020:2012 [157]), a validation/verification body (according to DIN EN ISO/IEC 17029:2020 [158]) or one of the certification bodies (according to DIN EN ISO/IEC 17021-1:2015 [22], DIN EN ISO/IEC 17024:2012 [155], DIN EN ISO/IEC 17065:2013 [17]).

In this context, certification is by definition a “third party” activity. The other bodies (testing laboratory, inspection, validation or verification body) may well be recognized or accredited as internal or not fully organizationally independent, but nevertheless competent and impartial conformity assessment bodies.

Accreditation of a testing laboratory, inspection body, validation or verification body is recommended if the conformity assessment activities for a specific AI system are to be offered e.g. independently of a subsequent AI certification using DIN EN ISO/IEC 17065:2013 [17]. Tests, inspections and validations ensure the quality of an AI system as a product according to Art. 6 para. 1 in conjunction with Annex II of the EU AI Act.

For specific AI systems or for such products (Art. 6 para. 1 in conjunction with Annex II of the EU AI Act [4]), test results from accredited testing laboratories according to DIN EN ISO/IEC 17025:2018 [156], inspection results from accredited inspection bodies according to DIN EN ISO/IEC 17020:2012 [157] or validations according to DIN EN ISO/IEC 17029:2020 [158] can potentially be generated independently of AI certification. Nevertheless, testing, inspection or validation results for specified AI systems are required to enable AI product certification. Here, a certification body according to DIN EN ISO/IEC 17065:2013 [17] may determine these itself or consider adopting previous test, inspection or validation results.

Accreditation is an important tool that ensures confidence in the comparability of the work of conformity assessment bodies and thus actively contributes to the removal of technical barriers to trade between countries. The basis of the practical working procedures for accreditation bodies is International Standard DIN EN ISO/IEC 17011:2018 [159]. It specifies the requirements for the competence, uniform operation and impartiality of accreditation bodies that assess and accredit conformity assessment bodies (see Figure 29).

The legal requirements regarding conformity assessment and accreditation are – as in the product world – specified in technical standards. In the field of conformity assessment and accreditation, there are clear distinctions to the level that is assessed. This results in a level system, which can be found in the documents EA-1/06 A-AB:2022 [170] and IAF PR4:2015 [171].

The system of accreditation and conformity assessment serves to secure the quality assurance process chain. Starting from Level 5, the object of conformity assessment is always considered, which was produced or created by a company. These are products, processes, services or people for whom certain requirements for specific qualifications are specified. The legal requirements specified in standards must accordingly be complied with by the companies. They must declare conformity with these requirements and – also for AI systems

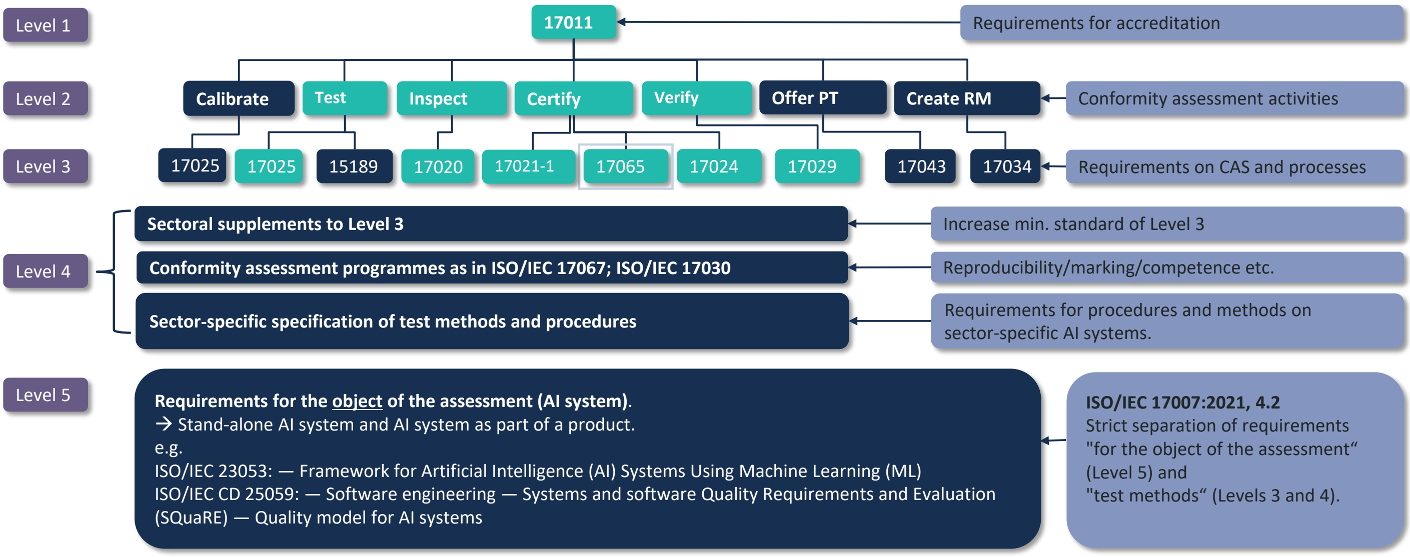

Level 1
17011
Requirements for accreditation
Level 2
Calibrate
Test
Inspect
Certify
Verify
Offer PT
Create RM
Conformity assessment activities
Level 3
17025
17025
15189
17020
17021-1
17065
17024
17029
17043
17034
Requirements on CAS and processes
Level 4
Sectoral supplements to Level 3
Increase min. standard of Level 3
Conformity assessment programmes as in ISO/IEC 17067; ISO/IEC 17030
Reproducibility/marking/competence etc.
Sector-specific specification of test methods and procedures
Requirements for procedures and methods on sector-specific AI systems.
Level 5
Requirements for the object of the assessment (AI system).
→ Stand-alone AI system and AI system as part of a product.
e.g.
ISO/IEC 23053: — Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)
ISO/IEC CD 25059: — Software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Quality model for AI systems
ISO/IEC 17007:2021, 4.2
Strict separation of requirements "for the object of the assessment“ (Level 5) and "test methods“ (Levels 3 and 4).

Figure 29: Classification of conformity assessment procedures in international level structure (Source: DAkkS)

as products according to Art. 6 para. 1 in conjunction with Annex II of the EU AI Act – demonstrate this conformity. In the case of a conformity assessment by a conformity assessment body, the normative requirements of Level 3 must be complied with by this body in order to carry out an evaluation of the object of the assessment (Level 5) on a scientific basis, to achieve comparable results and thus to be able to confirm the declared conformity of the manufacturer or the person placing the product on the market (Chapter 4.3.1.1).

Dimensions of AI trustworthiness (test dimensions)

DATA QUALITY AND DATA MANAGEMENT

The quality and trustworthiness of an AI application is closely linked to data quality. Data quality requirements include, for example, correct data annotation or trusted and relevant data sources [81]. Sufficient data quality is an important foundation for many of the other dimensions, such as a measure to ensure fairness or to achieve sufficient performance of an AI system. Closely related to this are the requirements for sensible data management that maps these quality requirements or governs data access, for example. The EU draft regulation [4] also formulates data quality and management requirements for high-risk systems.

BIAS, FAIRNESS AND AVOIDANCE OF UNDESIRED DISCRIMINATION

A fundamental requirement for the trustworthiness of an AI system is the avoidance of undesirable discrimination [154]. This requirement is intended to ensure that unjustified unequal treatment of individuals or groups in comparison with other groups is prevented [63]. Causes of undesirable discriminatory model behaviour often result from historical data that are unbalanced or that exhibit bias with respect to a particular group. Based on evaluation measures [see Working Group Basic topics], the nondiscrimination of an AI application can be quantified, where bias and undesired discrimination can be measured in the training data and in the output of the model.

AUTONOMY AND CONTROL

The ability to autonomously learn models and training parameters from data results in a degree of autonomy for certain AI applications. Depending on the context and criticality of an application, the autonomy of the AI application creates a conflict with the human autonomy of the users and those affected. To safeguard the primacy of human action, this conflict must be checked by an appropriate degree of autonomy between the AI application and user autonomy. At the same time, the dimension of autonomy and control also covers the requirement that users and stakeholders be adequately informed and empowered to interact with an AI application [120].

EXPLAINABILITY, INTERPRETABILITY AND TRANSPARENCY

Transparency includes various aspects such as interpretability, explainability, traceability or reproducibility of the results and functionality of an AI application. The reproducibility of results of a system is a minimum requirement for the traceability of results. While the interpretability of a system implies that the system as a whole is comprehensible [120], explainability merely means that it is comprehensible which factors have led to the result [16]. Transparency must be ensured in an appropriate manner and to an appropriate degree, so that it is accessible and adapted to the respective user [4].

PERFORMANCE, CAPABILITY, RELIABILITY, ROBUSTNESS, COMPLETENESS

To make an AI trustworthy, users must be able to rely on the system. From a technical point of view, the reliability of a system includes various aspects such as the correctness of the outputs as a rule, the estimation of the uncertainty of the results, or the robustness against attacks, errors and unexpected situations [120]. Performance metrics allow a measurable, qualitative and quantitative assessment of the system [16]. Even though Art. 15 of the EU regulation sets requirements for the reliability of high-risk systems, for example, the translation of the requirements into quantitative measures and target values remains open so far and requires specific knowledge of domains and applications.

SAFETY, SECURITY AND PRIVACY

Another dimension for testing and certifying trustworthiness is security/safety with the topics safety, information security, privacy, security and reliability. These are presented in detail in Chapter 4.2 "Safety/security".

4.3.2.2 Operationalization of AI tests

The goal of this chapter is to present the complex interrelationships and responsibilities for AI systems and the resulting implications for AI test methods that go beyond previous considerations of AI conformity assessments. According to preliminary considerations so far, test methods for the trustworthiness of AI systems should address three observations: complex AI supply chains, the hybrid nature of many AI systems, and embedding in technical systems.

AI supply chains

AI applications and AI services can stand as independent modules in delivery and performance relationships to other components of an (information) technical system that are relevant for the evaluation of the overall system. For example, an AI-based solution for credit card fraud detection can be built as a composite AI system consisting of three modules. The credit card operator provides transaction data that enters an AI service provider's learning system as raw datasets. The learning system produces rules for an expert system located at a financial institution that makes online "just in time" fraud detection recommendations. In this case, three actors are involved who independently realize parts of the overall system:

→ The credit card operator is responsible for the quality of the training and test datasets (B2B relationship in the overall system).
→ The AI service provider is responsible for the quality of the learned rules (B2B relationship in the overall system).
→ The financial institution is responsible to the end customer for the quality of the entire fraud detection process (B2C relationship in the overall system).

As a rule, these supply and service relationships with regard to the AI requirements of the components cannot be mapped in sufficient detail contractually; instead, the overall system, i.e., each of the modules included, must be examined and the individual test results combined to form a conformity assessment of the fraud detection system. The situation is shown schematically in Figure 30 with the integration of cloud service providers.

Hybrid AI systems

AI systems can be technologically hybrid, i.e., they consist of several modules with different AI technologies. A system for the recognition of spoken language, for example, consists at least of

→ an analogue-digital converter (microphone) to generate a speech spectogram by means of Fourier transformation, from which phonemes can be digitized via the frequency, time and intensity of the analogue signals.
→ Phonemes can vary depending on the speaker, accent, age, gender, or position in the word. Special AI technologies can be used to recognize words and sentences. Special Markov models are suitable as static models.
→ However, these are not flexible enough, especially in the event of a fault, so they are supported and safeguarded by further methods, e.g. by special neural networks.

The resulting products function very satisfactorily, which is confirmed in the technology-determined everyday life. However, it must be possible to verify the technologies on the basis of existing or AI test methods which are still to be developed, in such a way that a qualified statement can be made about the trustworthiness of the overall system. In such cases, the linkage, relevant to testing, with other questions becomes interesting, e.g., which technologies run on the client, which in the edge? In other words: What is the trustworthiness of the AI system based on its parts?

Embedded AI systems

AI applications, AI services, and AI modules in complex systems are specialized information technologies whose individual testing must provide usable results for existing test methods of the overall systems in which the AI can be embedded. The results of AI tests must be able to feed into the results of higher-level tests based on existing testing and approval procedures. For AI conformity assessments, the possibility to be included in existing certification procedures based on DIN EN ISO/IEC 17065:2013 [17] as partial assessments in the sense of an adoption of the results (according to DIN EN ISO/IEC 17065:2013 [17] see 9.6) must be taken advantage of.

These three observations lead to a whole series of interrelationships with in-progress and existing standards. The standards relevant to this chapter are listed below in the existing framework of conformity assessments. In view of the above observations, adaptations of this framework to the specific character of AI systems need to be considered at the same time.

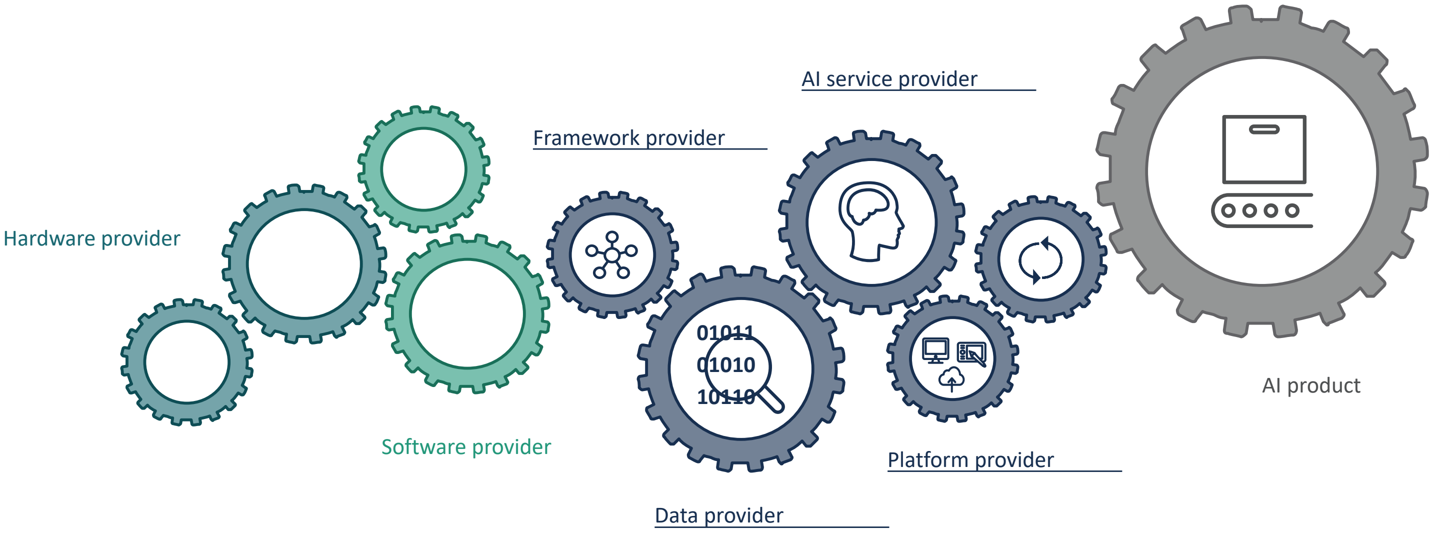


Figure 30: Actors in a cloud-based AI supply chain (Source: PwC)

Document	Title
Level 5 standards (Requirements on the object of conformity assessment)	
ISO/IEC 5259-2 [41]	Artificial intelligence – Data quality for analytics and machine learning (ML) – Part 2: Data quality measures
ISO/IEC 5259-5 [44]	Artificial intelligence – Data quality for analytics and machine learning (ML) – Part 5: Data quality governance
ISO/IEC TR 5469 [33]	Artificial intelligence – Functional safety and AI systems
ISO/IEC TS 5471 [34]	Artificial intelligence – Quality evaluation guidelines for AI systems
ISO/IEC 24029-2:2022 [92]	Artificial intelligence (AI) – Assessment of the robustness of neural networks – Part 2: Methodology for the use of formal methods
ISO/IEC TR 24029-1:2021 [91]	Artificial Intelligence (AI) – Assessment of the robustness of neural networks – Part 1: Overview

Document	Title
ISO/IEC 22989:2022 [16]	Artificial intelligence – Concepts and terminology
DIN SPEC 92001-2:2020 [240]	Artificial Intelligence – Life Cycle Processes and Quality Requirements – Part 2: Robustness
ISO/IEC 5259-1	Data quality for analytics and ML – Part 1: Overview, terminology, and examples
ISO/IEC 5259-3 [42]	Data quality for analytics and ML – Part 3: Data Quality Management Requirements and Guidelines
ISO/IEC 5259-4 [43]	Data quality for analytics and ML – Part 4: Data quality process framework
ISO/IEC TS 8200 [37]	Information technology – Artificial intelligence – Controllability of automated artificial intelligence systems
ISO/IEC 8183 [45]	Information technology – Artificial intelligence – Data life cycle framework
ISO/IEC 42001 [27]	Information Technology – Artificial intelligence – Management system
ISO/IEC TS 6254 [36]	Information technology – Artificial intelligence – Objectives and approaches for explainability of ML models and AI systems
ISO/IEC TR 29119-11 [132]	Information technology – Artificial intelligence – Testing for AI systems – Part 11:
ISO/IEC TS 12791 [38]	Information technology – Artificial intelligence – Treatment of unwanted bias in classification and regression machine learning tasks
ISO/IEC 24668	Information technology – Artificial intelligence – Process management framework for Big data analytics
ISO/IEC 5338 [30]	Information technology – Artificial intelligence – AI system life cycle processes
ISO/IEC TS 4213 [29]	Information technology – Artificial Intelligence – Assessment of machine learning classification performance
ISO/IEC 5339 [31]	Information Technology – Artificial Intelligence – Guidelines for AI Applications
ISO/IEC 5394 [149]	Information Technology – Criteria for concept systems
ISO/IEC 5392 [32]	Information technology – Artificial intelligence – Reference Architecture of Knowledge Engineering
ISO/IEC 23894:2022 [25]	Information Technology – Artificial Intelligence – Risk Management
ISO/IEC TS 24462 [150]	Ontology for ICT Trustworthiness Assessment
ISO 24089 [151]	Road vehicles – Software update engineering
ISO/IEC 23053:2022 [24]	Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)
ISO/IEC 27034-1:2011 [122]	Information technology – Security techniques – Application security – Part 1: Overview and concepts

Document	Title
ISO/IEC 27034-2:2015 [123]	Information technology – Security techniques – Application security – Part 2: Organization normative framework
ISO/IEC 27034-3:2018 [124]	Information technology – Application security – Part 3: Application security management process
ISO/IEC 27034-5:2017 [125]	Information technology – Security techniques – Application security – Part 5: Protocols and application security controls data structure
ISO/IEC 27034-6:2016 [126]	Information technology – Security techniques – Application security – Part 6: Case studies
ISO/IEC 27034-7:2018 [127]	Information technology – Security techniques – Application security – Part 7: Assurance prediction framework
DIN EN ISO/IEC 29101:2022 [493]	Information technology – Security techniques – Privacy architecture framework
DIN EN ISO/IEC 29134:2020 [134]	Information technology – Security techniques – Guidelines for privacy impact assessment
DIN EN ISO/IEC 29147:2020 [494]	Information technology – Security techniques – Vulnerability disclosure
DIN EN ISO/IEC 29151:2022 [135]	Information technology – Security techniques – Code of practice for personally identifiable information protection
ETSI DGR SAI 002:2021 [497]	Securing Artificial Intelligence (SAI); Data Supply Chain Report
ETSI DGS SAI 003 [336]	Securing Artificial Intelligence (SAI); Security Testing of AI
DIN EN ISO/IEC 27001:2017	Information technology – Security techniques – Information security management systems – Requirements
DIN EN ISO/IEC 27002 [481]	Information security, cybersecurity and privacy protection – Information security controls
DIN EN ISO/IEC 27701:2021 [128]	Security techniques – Extension to ISO/IEC 27001 and ISO/IEC 27002 for privacy information management – Requirements and guidelines
ISO/IEC 25000:2014 [472]	Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – Guide to SQuaRE
ISO/IEC 25024:2015 [473]	Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – Measurement of data quality
ISO/IEC 25020:2019 [474]	Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – Quality measurement framework
ISO/IEC 25010:2011 [152]	Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – System and software quality models
ISO/IEC 25021:2012 [475]	Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – Quality measure elements

Document	Title
ISO/IEC 25012:2008 [463]	Software engineering – Software product Quality Requirements and Evaluation (SQuaRE) – Data quality model
DIN ISO 31000:2018 [160]	Risk management – Guidelines
ISO/SAE 21434:2021 [324]	Road vehicles – Cybersecurity engineering
ISO 26262 series [455]	Road vehicles – Functional safety
ISO/IEC TR 24027:2021 [436]	Information technology – Artificial intelligence (AI) – Bias in AI systems and AI aided decision making
ISO/IEC TR 24372:2021 [437]	Information technology – Artificial intelligence (AI) – Overview of computational approaches for AI systems
ISO/IEC TR 24030:2021 [293]	Information technology – Artificial intelligence (AI) – Use cases
ISO/IEC 38507:2022 [26]	Information technology – Governance of IT – Governance implications of the use of artificial intelligence by organizations
ISO/IEC TR 24368:2022	Information technology – Artificial intelligence – Overview of ethical and societal concerns
ISO/IEC TR 24028:2020 [28]	Information technology – Artificial intelligence – Overview of trustworthiness in artificial intelligence
ISO/IEC 25059:2022 [35]	Software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – Quality model for AI-based systems
Level 4 Standards (Standards on designations, specification of test methods)	
DIN EN ISO/IEC 17050-1:2010 [489]	Conformity assessment – Supplier's declaration of conformity – Part 1: General requirements
DIN EN ISO/IEC 17050-2:2005 [490]	Conformity assessment – Supplier's declaration of conformity – Part 2: Supporting documentation
DIN EN ISO/IEC 17030:2021 [486]	Conformity assessment – General requirements for third-party marks of conformity
DIN EN ISO/IEC 15408-1:2020 [445]	Information technology – Security techniques – Evaluation criteria for IT security – Part 1: Introduction and general model
DIN EN ISO/IEC 15408-2:2020 [446]	Information technology – Security techniques – Evaluation criteria for IT security – Part 2: Security functional components
DIN EN ISO/IEC 15408-3:2021 [447]	Information technology – Security techniques – Evaluation criteria for IT security – Part 3: Security assurance components
ISO/IEC 15408-4:2022 [448]	Information technology – Security techniques – Evaluation criteria for IT security – Part 4: Framework for the specification of evaluation methods and activities

Document	Title
ISO/IEC 15408-5:2022 [449]	Information technology – Security techniques – Evaluation criteria for IT security – Part 5: Pre-defined packages of security requirements
DIN EN ISO/IEC 18045:2021 [75]	Information technology – Security techniques – Methodology for IT security evaluation
Level 3 Standards(Requirements for conformity assessment bodies)	
DIN EN ISO/IEC 17020:2012 [157]	Conformity assessment – Requirements for the operation of various types of bodies performing inspection
DIN EN ISO/IEC 17021-1:2015 [22]	Conformity assessment – Requirements for bodies providing audit and certification of management systems – Part 1: Requirements
DIN EN ISO/IEC 17021-3:2019 [485]	Conformity assessment – Requirements for bodies providing audit and certification of management systems – Part 3: Competence requirements for auditing and certification of quality management systems
DIN EN ISO/IEC 17024:2012 [155]	Conformity assessment – General requirements for bodies operating certification of persons
DIN EN ISO/IEC 17025:2018 [156]	General requirements for the competence of testing and calibration laboratories
DIN EN ISO/IEC 17029:2020 [158]	Conformity Assessment – General principles and requirements for validation and verification bodies
DIN EN ISO/IEC 17043:2022 [488]	Conformity assessment – General requirements for the competence of proficiency testing providers
DIN EN ISO/IEC 17065:2013 [17]	Conformity assessment – Requirements for bodies certifying products, processes and services
Level 1 Standard (Requirements for accreditation bodies)	
DIN EN ISO/IEC 17011:2018 [159]	Conformity assessment – Requirements for accreditation bodies accrediting conformity assessment bodies
Level 0 Standard (General principles of accreditation and conformity assessment)	
DIN EN ISO/IEC 17000:2020 [147]	Conformity assessment – Vocabulary and general principles

In the quality assurance and evaluation of information technologies, there are

→ test criteria for defining and describing system functionality,
→ criteria by which trust in the effectiveness of system functions can be assessed, and
→ criteria according to which the correctness of the test object with regard to the specifications of trustworthiness can be examined during commissioning and operation.

For the hybrid embedded AI systems and their supply and service relationships considered at the outset, all three types of criteria are required in a common test method. Such criteria-based testing and assessment of AI systems can be covered in an application-specific manner as part of a certification programme, and is referred to as evaluation.

Mapping vertical risks of overall systems into horizontal testing requirements for AI components

From the standpoint of evaluating trustworthiness of aspects of the test dimensions, there are two initial situations. Either the evaluation object (EO) is described in a concrete environment within a technical system, e.g., as a camera-based object recognition system in motor vehicles, or – and this is becoming increasingly common – the EO is available as an AI technological standard that is used as a "blank" and then individualized and adapted according to concrete deployment requirements, such as in an AI service of a financial service provider that processes transaction data as raw data and provides indicators for forecasting business transactions.

In both cases, a test will take the entire technical system into consideration. In each case, a risk analysis is performed on the basis of usage scenarios or supply and service relationships. Depending on the test dimension, different procedures can be applied (e.g., consideration of worst-case scenarios versus threat analyses). It is crucial that the hazards of the impact of the technical system on its environment and the hazards of the impact of the environment on the technical system are considered across all test dimensions on the basis of predefined exceptional situations, so that any (inter) dependencies that may arise between individual test aspects are identified and classified in the risk analysis process.
The term hazard is used in the context of risk analysis of AI systems for events that lead to undesired deviations of the specified behaviour of the overall technical system.

Risk analysis in this context refers to the complete process of assessing (identifying, estimating and evaluating) risks. However, according to the relevant ISO standards DIN ISO 31000:2018 [160] and ISO/IEC 27005:2018 [161], risk analysis refers to only one step in the risk assessment process that is required for risk treatment. The risk analysis for AI systems is based on ISO/IEC 23894:2022 [25] and consists of the creation of a hazard overview, i.e., a list of possible elementary hazards, and the identification of additional hazards that go beyond the elementary hazards and result from the specific deployment scenario, and a risk classification, i.e., an assessment of the risks after determining the frequency of occurrence and damage potential, and the classification into a risk category based on this. Risk treatment follows the risk analysis and consists of avoidance, reduction, transfer, and acceptance strategies, including the definition and testing of countermeasures. Measures taken to avoid or reduce AI-based risks include contractual agreements with AI service providers, software license agreements, and other quality assurance measures, e.g., through the use of testing tools, and more.

In German, the term "Risikoanalyse" (risk analysis) has become established for the complete process of risk assessment and risk treatment. In the evaluation of AI systems, however, the two steps must be kept separate.

The AI risk analysis process[82] performs the transfer of hazards from or to the technical system into risks that result in a set of requirements for the structure and functioning of the technical system. Depending on the test dimension, the concrete form of the specification can be taken from the relevant standards and specifications (see list above) or can, for example, follow the system decomposition according to DIN SPEC 92001-1:2019 [162]. In specific sectors, there may be a need to draw on additional standards, testing schemes, and technical control tools on a risk-based basis. Only in this way can the requirements appropriate to the test object be taken into account in the respective context. The specification usually contains the minimum requirements that can be derived from the risks, which are placed on a system component or which a component places on other system components.
For the AI modules contained in the technical system or for the AI components contained in the supply chain, a separate document must specify which requirements which AI module or which AI component expects or must fulfil with regard to which test dimensions.

[82] For a risk-based approach to evaluating AI systems, see also [120].

Through this step-by-step refinement, the risk analysis process finally extracts, at the level of AI modules and AI components, target objects with minimum requirements compliance with which is indispensable for the risks of the overall system. These requirements form the basis for the specification of the evaluation object in the system description. After testing has been performed, the results can be traced back analogously to the stepwise refinement and finally assigned to the risks at the level of the technical system or the distributed AI system. This multi-step refinement process and its tracing back with the test results of the AI components and AI modules is necessary to be able to develop the basis for cross-application AI certification required in the international standardization organizations.

Conformity assessments for AI systems derive minimum qualitative requirements from the AI risk analysis and mapping processes outlined above (see Figure 31). Such minimum requirements may address the operating environment of the AI system or relate to the development and specification process of the AI system itself. For example, in the context of developing a distributed AI system, an information transfer process should have been initiated, roles and tasks can be defined, and a structural analysis identifies key information about the entire system. This may result in other areas of consideration, such as:

→ When considering process and business risks, AI components can be explicitly considered and evaluated as a source of risk,

→ Certain risk-relevant parameters can be immediately included in the risk assessment, such as whether personal data is used, whether external data is used, and whether property damage or personal injury may result from the AI component,

→ Specific documents can be included to demonstrate test objectives, e.g. assurance cases as output of the assurance case method.

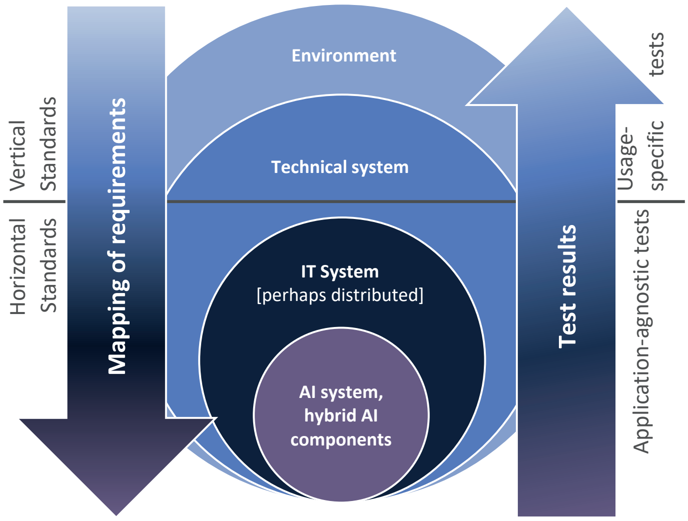


Figure 31: Step-by-step refinement of test requirements and referencing of test results (Source: BSI)

These minimum requirements and the guidance for documenting the mapping of risks into requirements for the AI modules and components are to be developed.

Principles of AI testing

The following are the basic principles of AI testing. The first step is to describe the test object and perform a risk analysis. Against the backdrop of determining the potential for harm to, for example, data, finance, fairness, and human mental as well as physical well-being, the application-specific description of AI systems is essential. Current standardization projects include approaches suitable for the concretization of risks, including the descriptions on risk management in the document ISO/IEC 42001 [27] (General description of an AI management system, for a presentation of the AI management system see also the study [120]). The basis of any evaluation is the description of the evaluation object (EO), i.e., the AI system whose trustworthiness is to be tested. An EO that is to be trusted must have certain properties. In order for a reasonable degree of confidence to be placed in the properties, they must themselves be described with sufficient precision. The accuracy of the description here depends on what AI technology(ies) the EO is using, for what purpose, in what way, and the depth of the trust to be placed in these properties. These details, representations and descriptions form a set of documents called test specifications. As a rule, each conformity test of an AI system requires its own test specifications. From the perspective of the EO, the test specifications address the questions:

→ What should be tested?

→ With which test depth should it be tested?

From this, a body responsible for the test can derive a specific test plan.

PROCEDURE MODEL

The first question is about the scope of the function of the EO, that is, its functionality. The second question is aimed at the trust that can be created by testing this functionality. The distinction between the functionality of a system and the trustworthiness required by test quality and test depth is one of the fundamental paradigms for criteria-based testing and evaluation of security properties of programmable IT systems – and thus also for AI systems. Criteria-based test methods first generate individual test plans tailored to the AI system using the test specifications. Functionality testing initially assigns test objectives to risks, which are then progressively refined. System functions are assigned to the test objectives at the level of the rough specification. One observation level lower, concrete measures are assigned to the functions – in the refined specification – which implement the functions. The test quality considers aspects of the effectiveness of the measures and the correctness of the implementation. Basically, this approach can be understood as a feedback waterfall model. For AI systems, both aspects – effectiveness and correctness – must be extended.

EFFECTIVENESS ANALYSIS

The effectiveness criteria to be developed as part of a test scheme should take into account the life cycle phases of the system and have different test focuses depending on the phase, such as ...

design:

- → analysis of the suitability of the mechanisms,
- → analysis of the interaction of the mechanisms,
- → analysis of the strength of the mechanisms,
- → analysis of the weakness of the design (for implemented mechanisms).

operation:

- → analysis of the test processes in the life cycle or for repeat tests (for test mechanisms).

Requirements for test tools can be derived from the properties of the effectiveness criteria. Test tools should provide all necessary information to interpret results appropriately. Such information should cover at least the following dimensions:

- → Scope and depth: What specific part of the AI system is being tested? What are the inputs and outputs of this part? Which and how much data is used to test the system?
- → Function assignment: Which functions are supported with the tool? What is a desirable outcome of the test? What is an undesirable outcome of the test?
- → Functioning of the test tool: The technical method used to test the AI system should be described. Limitations of the test method used should also be explicitly presented, as well as information on the stability and reproducibility of the test results.

CORRECTNESS

For correctness criteria, the approach of defining test criteria in stages, with each stage building on the next stage down, is appropriate. Such Evaluation Assurance Levels (EALs) are presented in Chapter 4.1.2.2. With the help of the EALs, the test quality and also the test depth are increased step by step. As part of the foundation for AI testing, a distinction between design and operational phases in the AI life cycle will need to be made at each stage. For each individual evaluation level, the evaluation criteria will then need to be further broken down into different phases. So far, the following phases seem to be relevant for correctness:

Design, development process:

1. requirements for the test specifications
2. architecture design
3. refined specification
4. implementation

Design, development environment:

1. approach
2. control processes
3. trustworthiness of developer

Operation:

1. specifications for the operation
2. delivery and configuration
3. start-up and operation
4. operational documentation
5. in-service testing
6. safeguarding of evidence
7. end of operations

Each phase will define the test measures and the documents to be provided at the start of the test and will specify the minimum requirements for the test results.

Quality infrastructure

This chapter argues for a universal certification process for AI systems based on existing standards and specifications, and on current international AI standardization activities. It was shown how such a method could be designed so that, on the one hand, it could be used for vertical AI standardization and, on the other hand, it could be linked to existing information technology testing and certification procedures. It was argued that such a method can provide directional implementation impulses for the implementation of AI regulation in Europe and at the same time achieve international market penetration. The argumentation is clearly for an AI certification program to be developed based on existing standards and specifications within a quality infrastructure that meets the following framework:

→ The certification programme is anchored internationally in two standards – "Trustworthy Artificial Intelligence Systems Evaluation Criteria" and "Trustworthy Artificial Intelligence Systems Evaluation Methodology".
→ The certification programme can be connected to existing IT testing infrastructures.
→ AI testers at conformity assessment bodies are promoted in special training and advanced training programmes (licensing, personal certification) within the framework of the tasks from the Standardization Roadmap AI on the basis of internationally developed quality requirements.
→ In conformity assessment, checks are made against applicable legal requirements and technical specifications – normative and ethical aspects are excluded.
→ The interfaces to AI management systems – especially to the AIMS – are clearly defined.
→ Certification and approval of AI test tools will be firmly anchored as a focal point in the above-mentioned sets of criteria.

4.3.2.3 Existing approaches and results

This chapter briefly presents projects and initiatives that have national and international significance in the context of testing and certification of AI systems.

ZERTIFIZIERTE KI

In the lighthouse project "ZERTIFIZIERTE KI" of the competence platform KI.NRW, a consortium of Fraunhofer IAIS, BSI, DIN and other research partners is developing test criteria, methods and tools for AI systems in order to make the quality of AI applications assessable by independent testers. Industrial needs are being taken into account through the active involvement of numerous associated companies and organizations representing various sectors such as telecommunications, banking, insurance, chemicals and trade. The results are being transferred to standardization.

A first result of the project is the "Guideline for the Design of Trustworthy Artificial Intelligence" [120], which provides developers with a guideline for systematically designing new AI applications in a trustworthy manner. It also guides testers in examining AI applications for trustworthiness in a structured manner. Here, the guide follows a four-step approach:

1. A comprehensive risk analysis along the dimensions of fairness, autonomy and control, transparency, reliability, security and data protection.
2. The establishment of objective targets, as measurable as possible, to make mitigation of the risks identified in 1 demonstrable.
3. A systematic listing of actions along the life cycle of an AI application to achieve the targets set in 2.
4. The creation of a stringent argument that the targets formulated in 2 have been achieved ("assurance argument for trustworthiness"), also taking into account AI-specific trade-offs, e.g., security vs. transparency.

For more information, go to the project's home page www.zertifizierte-ki.de.

A test standard for cloud-based AI[83]

Broad market access for tested AI in clouds can be ensured by performing (relatively low-cost) compliance tests at the level of the AI-based cloud service, addressing the effectiveness of measures against hazards or even risks to the AI within the cloud service. These tests, which are repeated regularly throughout the life cycle of the AI application, are based on three pillars:

1. Fundamental testing of the entire underlying cloud system from the infrastructure (IAAS) to the platform (PAAS) to the interfaces to the AI service (SAAS). There are already sets of criteria for such tests, such as the BSI's C5 criteria catalogue. In turn, where technical testing procedures are no longer applicable, this is based on the provider's personnel, organizational, institutional or spatial constraints if the residual risks beyond technical testing are not acceptable.
2. The in-depth technical testing of the provider's AI framework, which, after all, is initially offered in the same format for every customer. This assumes testing of various types, quality, and depth, up to and including certification processes, for individual AI technologies and processes. The test schemes for this must be developed and evaluated in the project. Beyond various test schemes, they can be designed in such a way that the content requirements are composed as a catalogue of criteria from an expandable set of building blocks. A general test methodology is developed for the test quality and test depth, which is incorporated into the standardization together with the criteria catalogue. Specific test

[83] See Chapter 6.6.

methodologies can then be derived from the general test methodology for different test schemes (compliance testing vs. certification).

3. The aforementioned elements must be transferred into an overarching standard from which a corresponding individualized test scheme can be derived for each of the cloud service providers, covering all relevant risks with personnel, organizational, technical and spatial dimensions in the sense of the instance of a risk-based management standard. The minimum standards (e.g., basic protection) of the BSI can prove to be useful building blocks beyond the technical quality features during instantiation.

For the overall success of the project and its implementation, it is therefore necessary to derive all the individual elements from a standard for trustworthy AI in cloud systems and, in the specific case, to assemble the building blocks from underlying schemas that are relevant for trustworthiness in the specific use case.

This lighthouse project of the Standardization Roadmap AI is being led and implemented by the BSI and is being accompanied by international standardization projects.

AI standards for medical diagnostic systems[84]

The project aims to develop test criteria and test methods for the use of AI in medical diagnosis and prognosis systems and to embed them in relevant standards in such a way that test standards for AI in medical technology can be established.

To do so, the following milestones must be met:

→ Develop extensible testing criteria for relevant AI technologies in medical technology based on existing standards and established specifications,

→ Evaluate these testing principles in pilot projects with deployed AI solutions as part of a continual improvement process,

→ Derive and develop reference architectures and testing profiles for the use cases considered below in the application domain and for AI technologies used with the goal of reducing testing efforts,

→ Standardize the developed test principles and criteria and classification on the basis of existing standards, and finally

→ Establish AI testing standards at international level.

This lighthouse project of the Standardization Roadmap AI is being led and implemented by the BSI.

ExamAI: Assurance Cases and Acceptance Test-Driven Development

In the project "ExamAI – Testing and Auditing of AI" funded by the German Federal Ministry of Labour and Social Affairs (BMAS), a combination of assurance cases and Acceptance-Test-Driven Development (ATDD) was proposed to support auditing and, in the long term, also certification of extra-functional requirements Assurance cases are structured arguments that explain why a system has been judged to be sufficiently good in terms of a specified property to be deployed. An assurance case starts with a quality claim, such as a system is fair. This claim is now divided into subclaims based on arguments (reasoning). Each argument can additionally be supplemented by contextual information (context) and assumptions. At the end, for each claim, there are evidences that prove that the respective claim is true. The concept originates from philosophy and is currently a common framework in safety engineering to argue on the basis of which arguments a system is considered sufficiently safe. The addition of ATDD requires that the assurance case be established before development begins. It is thus a concept of the test-first philosophy. A diverse group of stakeholders (project managers, developers, users, stakeholders, lawyers, ...) meets to develop theoretical scenarios in which the system could act against the property to be ensured. Based on this, possible countermeasures are developed, such as tests. At the end, the assurance case is made, arguing why the tests are considered sufficient.

ENISA

The EU's cybersecurity agency, ENISA, is tasked with contributing to a high common level of cybersecurity across Europe [118]. The Regulation applicable to ENISA is Regulation (EU) 2019/881 [163] of the European Parliament and of the Council of 17 April 2019 on ENISA and on cybersecurity certification of information and communication technology and repealing Regulation (EU) No 526/2013 (Cybersecurity Legislative Act) [164]. ENISA has published two papers on the subject of artificial intelligence:

– ENISA Report – Artificial Intelligence Cybersecurity Challenges with three topics: AI LIFE CYCLE; AI Assets; AI THREATS.

Content includes an overview of the AI cybersecurity ecosystem and its threat landscape, considering the AI life cycle. Five chapters present a generic reference model, details of

[84] See Chapter 6.6.

the AI ecosystem, a threat taxonomy linking relevant constituents and associated threats, and cybersecurity challenges for AI.

– ENISA Report – SECURING MACHINE LEARNING ALGORITHMS; December 2021 [119]
This report includes a taxonomy of ML algorithms, identification of relevant threats and vulnerabilities, and a list of security controls.

Building on ENISA's AI threat landscape mapping, this study focuses on cybersecurity threats specific to ML algorithms. It also suggests vulnerabilities related to the above threats and, most importantly, security controls and mitigation measures.

The adopted description of AI is a deliberate simplification of the state of the art with respect to this vast and complex discipline, with the intent not to define it precisely or comprehensively, but to contextualize the specific technique of machine learning pragmatically.

As a result, it was found that there is no clear strategy for applying a particular set of security controls to protect machine learning algorithms. The overall cybersecurity posture of organizations using machine learning algorithms can be improved by carefully selecting the controls developed for these algorithms.

Current activity: AI junior research group BAuA
The administrative agreement concluded between the BMAS and the BAuA describes a research strategy on the topic of "AI in a safe and healthy working environment". To implement the strategy, the BAuA has set up a junior research group for the next five years. The aim of the group is to provide answers to application-oriented questions about AI in the world of work within the framework of doctoral projects carried out in collaboration with relevant university institutes. Along the lines of the two areas of legislation on which the rules and regulations for ensuring safety and health at work in Germany are based, a distinction is made between two subject areas: occupational health and safety, and product safety. The challenges posed by the use of AI in the respective subject area are being explored in greater depth by two teams, one in Dortmund (occupational design measures) and one in Dresden (product safety).

Current activity: KI-LOK – A joint project on test methods for AI-based components in railroad operations
The design and operation of innovative vehicles in rail-based transport increasingly calls for the use of AI-based learning systems to improve the quality of transport services, increase resource efficiency and thus the sustainability of trains, and provide new functionalities. One of the greatest challenges in this context is the development of appropriate verification and validation methods, which in their entirety must meet the goals of data-based mobility, as well as the quality and safety requirements of rail transport, and must be suitable for proving the functional safety of AI systems. The aim of the project is to develop test methods for safeguarding and certifying AI-based technologies for safety-critical applications in railroad technology. The techniques and tools to be developed will be based on practical application examples in order to be practical. Therefore, based on two case studies – "Object recognition in the clearance gauge" and "Safe self-location as part of the vehicle odometry system" – the training and testing strategy for AI systems will be developed and made usable for industrial applications. The scope of the KI-LOK project is defined by the three cornerstones of approval processes, risk and hazard analysis, and analysis methods for AI. The results of the project form the basis for a tool-supported method for validating and verifying AI-based components in an industrial environment and also define a systematic framework for defining approval processes for AI-based applications in railroad operations. The KI-LOK project is supported and financed by the Federal Ministry of Economic Affairs and Climate Action (BMWK) within the framework of the funding guideline "New Vehicle and System Technologies" [165].

Current activity: Industrial Grade Machine Learning for Enterprises (IML4E)
Like to classical software, AI-based software must be implemented and validated according to end-user requirements and must meet the established quality attributes of classical software as well as a number of new quality attributes (e.g., interpretability, intelligent behaviour, non-discrimination, etc.). Its use must be technologically, socially and ethically acceptable and safe. All of this must be carefully planned, implemented, validated, and maintained throughout the software life cycle. Against this background, the IML4E project brings together companies from the main sectors of the German and European software industry to develop a European framework for the development, operation and maintenance of AI-based software, thereby ensuring the development of intelligent services and intelligent software on an industri-

al scale. The project focuses on providing industry-ready techniques, methods, and tools that are not currently freely available through open source solutions, and addresses established software development principles such as reuse, automation, and the tight integration of development and operations across the entire software life cycle, enabling German companies to integrate AI-based software into their development processes and products. The IML4E project is funded by the German Federal Ministry of Education and Research (BMBF) as part of the European ITE initiative [168].

4.3.3 Standardization needs

Need 03-01: Specification of formal requirements for "explainable" AI ("XAI") methods
Formulate concrete operationalizable/testable requirements for XAI methods.

What formal statements should be possible based on the results of an XAI method?

→ Concerning the training data?
→ Concerning the test date?
→ Concerning the model?
→ Concerning the relationship between input and output data (predictions)?
→ Concerning the relationship between model, input and output data?

Which practical consequences are to be securely derived from these statements? What added value of "reliability" is really to be created, and how can it be demonstrated?

A cross-sectoral requirement, which is also anchored in the draft AI Act, is that of "explainability," "interpretability," etc. However, there is a large gap between the legal/regulatory requirements and the concrete implementation of XAI. The XAI methods published in the literature do not yet close this gap, since the requirements for the methods are usually not specified concretely enough by the authors. Accordingly, the validation/verification of these methods often tends to be qualitative, subjective, and circular.

Formal criteria are necessary to specify which statements/practical consequences are correct and permissible based on the result of a given XAI method. Compliance with these criteria must be verified formally or empirically. This is the only way to avoid misinterpretations.

Need 03-02: Operationalization of the "explanatory quality" of XAI methods
Development of ground-truth reference datasets. The answers to questions to be provided by XAI methods are known for these data by construction and can therefore be matched with the result of XAI methods. The data can be generated by mathematical formation rules, physical simulation or manipulation of real data.

Development of appropriate metrics for "explanatory quality" of XAI method on ground-truth reference data (e.g., precision/recall, other metrics from signal detection theory).

Without sufficient verification of XAI methods themselves, it remains unclear what use they may have for the quality assurance of ML systems.

Need 03-03: Development of a standard with guidance documents for mapping the risks of a system into the functionality of AI components
AI systems are:

→ possibly hybrid,
→ possibly components of a technical systems,
→ possibly part of a distributed architecture on different platforms and in different infrastructures.

The risk analysis for the AI system is performed with a view to the entire technical system. Safety, security, etc. requirements for the parts and components of the AI system must be derived from this. This will have to be done taking into account the intended use and existing test specifications and framework conditions (ISO 26262 series [455], Machinery Directive, etc.). Thus, risks are mapped in whole or in part to testing requirements on the entire AI system or on parts of it. This mapping provides the anchor for embedding test results into existing test methods and evaluating them.

→ Contributions CEN/CLC JTC 21 & ISO SC 42 WG 3 "TAISEC" & "TAISEM"

Embedding AI testing into existing testing infrastructure.

Need 03-04: Development of functionality classes for AI technologies
Each AI system or product will have its own requirements regarding trustworthiness compliance. To meet these requirements, technical functions are available that the AI system either contains itself or that its environment provides, for example, for adversarial detection and defense, log evaluation,

or error detection and bridging. Appropriate trust in these functions is required, whether it is trust in the correctness of the specific functions (from both a development and an operational standpoint) or trust in the effectiveness of those functions. To be able to check both, a function must always be related to the AI technology that the AI system contains. Thus, different AI technologies have different possibilities, e.g., for error detection. These functionalities must be classified so that the functions can be easily assigned to the requirements. Thus, a building block of relevant functionality classes is needed.

→ Contributions CEN/CLC JTC 21 & ISO SC 42 WG 3 “TAISEC“ & “TAISEM“

Need 03-05: Development of tool criteria for testing AI systems
Tools for measuring properties of an AI system, such as performance, play a critical role in testing the system. The significance of the results of such measurements determines the significance of the entire test procedure. The appropriate test criteria and test methods are needed for the testing and certification of such tools. The emerging test procedures are part of the AI certification program to be developed in the above-mentioned standardization contributions.

→ Contributions CEN/CLC JTC 21 & ISO SC 42 WG 3 “TAISEC“ & “TAISEM“

Need 03-06: Development of interlocking standards for AI systems and necessary conformity assessment procedures
In order for conformity assessment procedures to be usable for AI systems, it is important that the applicable standards of the DIN EN ISO/IEC 17000:2020 series [147] (Level 3 standards) are observed. For specific requirements for particular evaluation tasks within the defined Level 3 conformity assessment activity, standards for AI systems differentiated by sectoral or technical requirements are to be developed at Level 4.

In addition, there is a need for standardization in the area of fundamentals, especially with regard to calibration and suitability testing providers (interlaboratory comparisons). Again, standards need to be developed at Level 4 that take into account the technical specifics and risks of AI systems.

It is particularly important to separate the standardization projects relating to the object of conformity assessment (AI system/organization in the sense of manufacturer or distributor) (Level 5) from the standardization projects relating to conformity assessment (Levels 4 and 3).

Only in this way is it possible to correctly assign the individual roles and responsibilities with regard to manufacturers, distributors, users and conformity assessment bodies.

Only through clear specifications of qualifications and clear requirements can test methods be developed that must be measured in their evaluation quality by interlaboratory comparisons. First, the requirements for the object must be known before it can be determined how these can be verified.

→ Contributions CEN/CENELEC JTC 21 WG 2 “Conformity Assessment“

Need 03-07: Development of qualification criteria for testers and certifiers of cybersecurity and privacy for AI
Development of a standard with criteria for the qualification of testers, auditors and certifiers for cybersecurity and privacy in AI, taking into account existing standards from the DIN EN ISO/IEC 27000 series [131].

Currently, there are established testing and certification procedures for the qualification of experts for the testing and certification of cybersecurity and privacy, but not yet for AI. These are also necessary.

Need 03-08: Networking of all actors
When developing standards, it is important to involve all stakeholders and interested parties, in particular authorities according to Art. 5 and Art. 7 of Regulation (EU) No. 1025/2012 [169], and to ensure the networking of experts from all required fields.

Standardization projects which provide for methods, procedures or processes, e.g. which provide for conformity assessment (e.g. as testing) of requirements, must be staffed with a broad field of experts from the area of conformity assessment bodies and accreditation bodies.

In this context, a common understanding of the necessary interaction of the various levels (metrology, conformity assessment, accreditation, manufacturing, placing on the market and use) must also be established within standardization work, in order to develop suitable and interlocking standards for the subject matter (e.g. AI system) and for conformity assessment (e.g. in the context of a test).

In standardization, there is no overarching understanding of how, in practice, standard requirements for the object and standard requirements for testing processes interrelate. In the future, a better attempt should be made to highlight this at the beginning of a standardization project in order to have better coordinated standardization projects. The better the mutual understanding, the easier the practical implementation.

In Chapter 4.3 it is made clear that the understanding of “testing and certification” is perceived differently, depending on the field of application and professional context. There is a legally regulated system in the EU that ensures the quality of products, processes, services and services: the quality infrastructure.

Need 03-09: Definition of control points
Based on the AI life cycle, individual test points at which a conformity assessment (Level 4 and 3) must take place are to be defined with a minimum set of evaluation activities in order to be able to assess and confirm conformity with the legal requirements defined in legislative projects such as the European AI Act or the Canadian Artificial Intelligence and Data Act [170].

A clear role definition is necessary at the level of the AI developers/manufacturers/distributors, as well as at the level of the conformity assessment bodies and accreditation bodies.

After a clearer role structure is established, it is then important to work out which roles (from Level 5 or Level 3) need to be integrated into the development, evaluation, deployment and decommissioning of the AI system at which point in the AI life cycle to meet regulatory requirements.

An improved networking of companies developing and/or placing AI systems on the market with conformity assessment bodies (first, second and third party).

The Working Group Testing and Certification ranked the identified needs according to the urgency of their implementation. Figure 32 shows the urgency of implementation, categorized according to the target groups of standardization, research and politics.

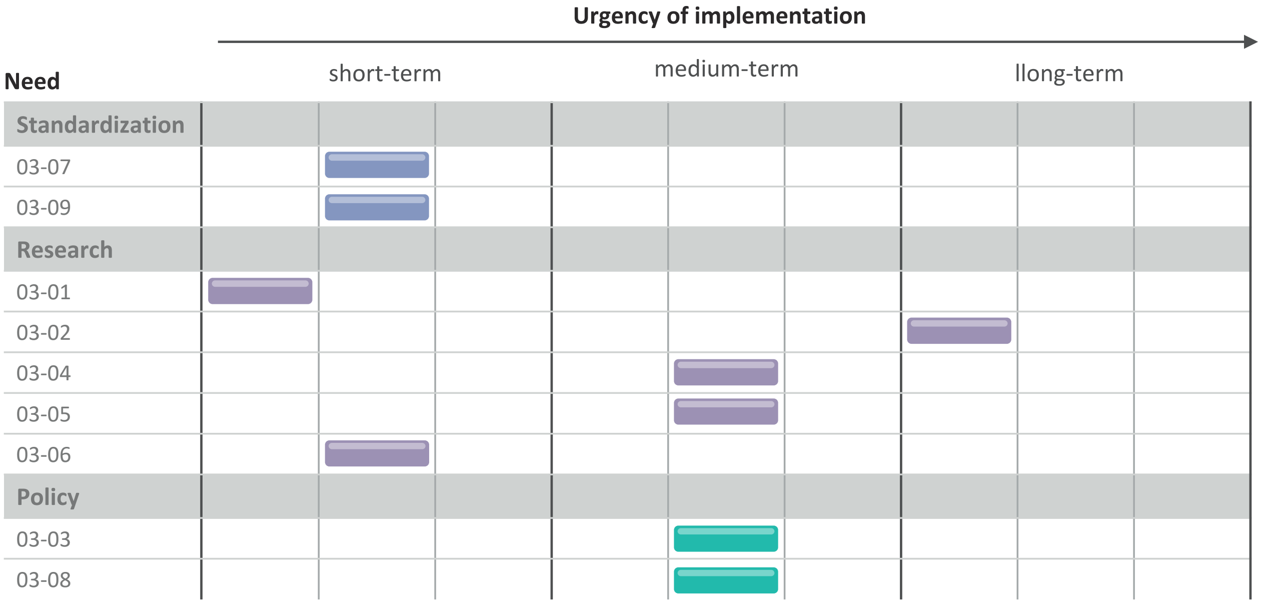


Figure 32: Prioritization of needs for the key topic Testing and certification (Source: Working Group Testing and Certification)

4.4 Sociotechnical systems

The topic of sociotechnical systems is considered in a separate chapter for the first time in this second edition of the Standardization Roadmap AI. Important preliminary work can already be found in the first edition in the chapter Ethics/Responsible AI with the aim to place value-oriented requirements on IT systems and to design and implement solutions that put humans at the centre. Moreover, ethical guidelines for algorithmic decision systems have been and are being discussed in different contexts (cf. [67], [173]). How to succeed in operationalizing the underlying ethical values in order to implement this requirement in concrete terms is the focus of this chapter.

4.4.1 Status quo

4.4.1.1 Classification of the sociotechnical system in the AI context

Sociotechnical systems include the subsystems humans and technology, which are interrelated and interact or should interact with each other (see [174], [175], [176]). In this context, AI technology is related to humans, the organizational environment, and society as a whole. Therefore, important issues include the integration of technology into societal subsystems, human-technology interaction [177], and organizational development [178].

Sociotechnical design of IT systems requires that they can support (work) tasks of humans in different roles and in the context of use, i.e., make them accessible to humans, for example, via ergonomically designed interfaces (e.g., displays and control elements) (e.g., [179], DIN EN 614 series [180], [181], [182] and DIN EN 894 series [515]). The user-centred or human-centred approach [183] puts people in focus. The basic principle is based on identifying people's needs, analyzing them, and using them to design products (AI) that help users complete their tasks effectively, efficiently, and satisfactorily.

Technology – Humans – Organization – Society

The introduction of AI applications into existing as well as new (work) processes enables the generation of positive potentials, but is also associated with challenges regarding the governance of these. AI should be seen as "a new class of agents in the organization" [184], which makes the term much broader than simply understanding it as a technical tool. This requires an understanding of how AI works and implies the organizational or process integration of AI applications [185], [186], [184], [187]. Many of the points and issues listed here for AI applications apply equally to "classical" algorithms.

In the interaction between humans and AI, degrees of autonomy can be distinguished (see e.g. [188]). These depend on how the interaction is designed [189]:

→ For example, AI can only execute something if the human confirms it first.

→ AI is more autonomous when it acts on its own, but humans can veto it.

→ AI could also act autonomously and inform humans only when they consciously ask for it.

→ Finally, AI could act without involving humans.

In the past, design concepts in ergonomics/human factors (EHF) (including sociotechnical design) mainly referred to static technical systems (e.g. interface design to static and stationary machine). Not only, but also through AI (as a dynamic system in terms of content and time with cause-effect relationships that can no longer be documented), the EHF design concept must be extended so that dynamics of interfaces, modes of operation and effects are also designed to suit humans.

The way we work is changing with the introduction of AI applications, and so are the demands on workers. Human attributes such as empathy or emotional dimensions will stand out in skill needs [190], [191], [192], [193]. In the context of AI, humans play different roles: Humans commission, develop, revise, and use AI and its results for their own purposes or on behalf of others (e.g. [194]). Not all humans use AI applications to the same extent and therefore they need corresponding competencies in the same breadth and depth (cf. [190]). AI, in turn, affects humans and their behaviour, so design must incorporate human performance requirements in interface, function, and effect [195]. Similar to the concept of communication [196], according to which humans cannot "not communicate", humans cannot "not interact with AI", as long as they are affected by it (e.g. clients, users, those impacted by effects). This makes sociotechnical design of AI technology necessary for its objective, functioning and effect in an overall system as well as for the task, interaction and information interfaces of human-technology interaction.

The concept of sociotechnical system design explicitly postulates the need to optimize the use of technology and the organization together ("joint optimization") (cf. [197] or [198], [199]). The organization thus represents the framework and, moreover, a central (in the best case a social partnership) level of regulation of the relationship between humans and technology and, at the same time, is conditioned by it in its internal structure. The description of the work task is central [199]. In turn, according to Ulich et al, the decisive factors include "the company's objectives, company strategy, company organization, market position, products and production conditions, personnel structure, use of technology, quality management, innovation behaviour, wage system, working time models, type of employee representation and negotiation processes, and the sociotechnical history of the company." Such an organization is in turn integrated into an environment (state and society, European and international agreements, standards and legislation).

Not only the organization and the human being, but also the societal perspective plays an important role in sociotechnical systems. In the respective application context, "society" is characterized by very different actors and values. Configurations of the subsystems human and technology come together at the interface "society". This can also perpetuate power and inequality relations or discriminatory patterns. Technologies aimed at automating human intelligence are not objective or neutral and may contribute to reinforcing racism and other phenomena of social inequality [200].

Humans and machines mutually influence and change each other in the usage process [176]. New developments such as machine learning illustrate this when software programs react dynamically and "adaptively" to their users [201]. This understanding of humans and machines challenges the previous conception of autonomous and strictly separable entities: It is only through the mutual adoption of, for example, linguistic rules that the human and machine actors are granted their ability to act and that an understanding arises that is collaborative in interaction [176].

The design of sociotechnical systems is based on the HTO concept, which assumes that the subsystems of human (H), technology (T) and organization (O) are linked by the work task and interact with each other [199]. In this context, not only the three subsystems themselves must be considered, but attention must also be paid to the interfaces human/technology, human/organization and technology/organization. There are target criteria for each of these interfaces. In addition, overarching target concepts can be formulated for sociotechnical design. For example, the concept "adaptivity, human-in-the-loop and human-centred technology" for the human/technology interface could be accompanied by concepts such as "holistic tasks and sense-making" (human/organization interface) and "decentralization" (organization/technology interface). From the starting point of a consistent sociotechnical human/technology/organization model, central questions of the introduction, use and impact assessment of AI can then be addressed in a more targeted manner [202], [203].

→ What data is AI associated with and for what purposes?
→ How does AI use affect human behaviours (e.g., autonomy, decision-making dilemmas, human behavioural adaptations)?
→ How does the AI application relate to human needs and expectations (e.g., the need to be able to assess and persuade one‘s counterpart)?
→ What are the systemic consequences of AI use within the system, for its subsystems, and also for the system environment and society (e.g., simple tasks become automated; difficult tasks become more difficult; security risks change as users adapt their behaviour to the automated technology)?

The special aspect of the sociotechnical perspective
As a proven thought model, the sociotechnical perspective connects earlier stages of industrialization with their human-technology interactions and the digital transformation in a connectable way. The use of AI in the work process requires human interaction with AI, which can be described as a work task. Technical, organizational and qualification elements interact in the work task [204], [205]; its hierarchical and sequential completeness (in the sense of the psychological action regulation theory, e.g. [206]) can be used as an assessment standard for the quality of the work (according to [207]).

The opportunities and risks of AI do not depend solely on the technology and its development, but on the context of its application. The sociotechnical perspective presents this context, facilitates operationalization, and is also the appropriate multi-perspective "counterweight" to a purely technology-centric view of AI. At the same time, this approach offers innovation potential because it is able to turn those affected into participants and offers a model of subsidiary (refined) regulation, e.g. at the company level.

Just as humans systematically make decision errors [208], "bias" effects or decision errors regarding fairness can also arise in the development and use of AI. "Bias" stands for undesirable distortions, which may already arise during the collection of the datasets themselves or due to the selection or type of processing. Last but not least, biases go back to design decisions (e.g., database and logic) and the underlying presuppositions of problem construction. With the use of AI, challenges also arise with regard to accountability and fairness when AI is used in cognitive applications (see practical example of processing applications). The undesirable effects around bias or fairness [209] refer to uncertainty about the consequences of AI application. For the decision-maker, risk refers to the occurrence of one or more known environmental conditions with an empirically determined probability of occurrence (e.g. tomorrow it will rain and the probability of occurrence is 70%). That means: Risk is quantifiable and thus potentially controllable. Uncertainty differs from risk in that neither the possible environmental conditions nor the possible probability of occurrence are known (e.g., the Covid 19 pandemic outbreak and subsequent effects)[85]. Existing algorithms for risk situations and estimation weigh risk in terms of probabilities of occurrence and desired optimization levels [208]. In addition to risk and uncertainty, there are other factors to consider (e.g., perceived process control [210] or decision depth of algorithms [211], etc.). Overall, it can be stated that human perception plays a crucial role in the analysis, design, and evaluation of AI systems.

Aspects of social sustainability in the sociotechnical context

What is understood normatively by sustainability is often negotiated and decided in parliaments and in national and international bodies and, if necessary, also implemented in law. When designing AI applications, it must therefore be ensured that they meet sustainability criteria. This results in the requirement for the AI system to be parameterizable with regard to quantitative targets from sustainability specifications.

In the sociotechnical context, aspects of social sustainability in particular must be considered. “With regard to the development, use and deployment of AI systems, sustainability means above all that human dignity is respected, no people are excluded, disadvantaged or discriminated against, and human autonomy and freedom of action must not be restricted by AI systems. In a broader perspective on sustainability, social sustainability also means that, in addition to physical integrity and decent living conditions, the ability to think, reason, and act in a human way should not be restricted. This already shows that a comprehensive understanding of social sustainability has very far-reaching consequences for the design of AI systems.” [223] At the same time, the wide range of objectives and aspects to be taken into account makes it clear that laws and standards reach their limits and cannot regulate every detail. Subsidiary negotiation systems, e.g. at the operational level, as well as individual decision-making rights are necessary. The sustainability of AI systems is ultimately negotiated and decided using different indicator systems and rules at the HTO level (HTO: humans, technology and organization).

AI systems can harm individuals and groups of individuals, which Muhammad (2022) [224] assigns to different types:

→ Thus, there are „allocation errors“ in that the system withholds or unfairly provides opportunities, resources, or information. One example here is discrimination in job application procedures, but also unequal treatment of people with and without Internet access [225].

→ Another category is „service quality errors“ where the system does not perform similarly for all groups.

→ A „representation error“ occurs when the development or use of a system over- or under-represents individual groups. As an example, there could be a predominance of males in an image search for „CEO” [226], [227].

→ Furthermore, a possible „stereotype error“ is listed, in which the system reproduces and reinforces stereotypes, for example, by assigning stereotypical characteristics to all members of a group without reflection.

→ A „disparagement error“ occurs when the system becomes actively derogatory or insulting, such as the reach-optimizing behaviour of Microsoft‘s Twitter bot Tay.

→ Finally, a „process error“ is the behaviour of a system that makes decisions based on characteristics that should not be relevant to the task. An example of this is job application process management that devalues people with too much work experience than that needed [228].

The Artificial Intelligence Act (AI Act) of the European Union (EU)

The present draft of the EU AI regulation addresses the sociotechnical perspective: “Artificial intelligence (AI) should be a tool for humans and a force for good in society, with the ultimate goal of enhancing human well-being. The European concept for artificial intelligence focuses on excellence and trust; it aims to foster research and industrial capabilities while ensuring security and fundamental rights.” (European Commission: Draft on the “Implementation plan of standardization requirements by the European standardization organizations”). “Among other things, the proposed AI Act introduces requirements for the placing on the market and commissioning of high-risk AI systems. These requirements

[85] https://wirtschaftslexikon.gabler.de/definition/risiko-44896/version-268200

address risk management, data quality and governance, technical documentation, record keeping, transparency and provision of information to users, human oversight, accuracy, robustness, and cybersecurity." This characterizes the draft in terms of defining the protection targets. The regulations themselves, on the other hand, relate more to the AI product and less to its application within the framework of (work) processes. Therefore, the draft only partially meets the requirements of a sociotechnical consideration.

The requirements to provide transparency and information for users and to ensure human oversight can only be fulfilled if the AI system is understood as a sociotechnical system and if humans are considered as part of that system. For this reason, when dealing with AI, it is of utmost necessity to clearly define the system boundaries of the sociotechnical AI system, to consider the interaction of its system elements, and, most importantly, to assess and design the interrelationship of the technical AI components with human behaviour. Because some systems continue to learn as they are used, a one-time test and optimization at a given point in time is not sufficient for the lifetime of a system (see Chapter 4.4.2.4).

Practical examples

Concrete scenarios are being discussed in the public debate: For example, regarding the automated selection of job application documents: A process developed by Amazon back in 2014 made headlines because it structurally disadvantaged women. According to Reuters (2018) [212], the algorithm had been trained with the datasets of accepted applicants – however, in the ten years used as a basis, mainly men had been hired, so that the algorithm concluded that applications from men were to be preferred. However, similar decision patterns have also been demonstrated for other personnel selection software solutions (cf. e.g. [213]). Training data or learning methods that themselves already contain a bias can lead to a "wrong" result with a so-called "bias". Similarly, the development of a reach-optimizing Twitter bot that learned to spread radical right-wing ideas within a very short time failed. Another current debate is about autonomous driving cars and their "decisions" in dilemma situations ([214], [215]). These problems bring to the fore the consequences of significant decisions regarding the data basis (e.g., sexist "bias" already in the training data), use and influenceability (e.g., learning in real time from unfiltered data), and interdependencies (e.g., legal consequences, user acceptance, etc.) in the development of AI-based algorithms – and justify the need for careful consideration of the solution to be selected and the choice of training data.

4.4.1.2 Interfaces to areas that cannot be standardized

European and national laws, regulations and rules take precedence over national and international standardization work. In Germany, for example, this applies to occupational health and safety and data protection.

According to Article 9 (3) of the German Grundgesetz (GG) (Basic Law), the social partners are responsible for "safeguarding and promoting working and economic conditions". Social partnership therefore extends to all areas of economic and social policy. In particular, this includes the regulation of all remuneration and other working conditions by collective agreements (collective bargaining autonomy). Accordingly, social partnership tasks belong to the area that does not relate to standardization. Standardization can at best supplement or give concrete detail to the legal framework here.

Health and safety rules and regulations

For the design of work systems, the development and consideration of standards and occupational science findings is not sufficient. Requirements for workplaces and work equipment are regulated by law at national and European level with regard to occupational safety and health, among other things. A fundamental distinction must be made here between legal requirements relating on the one hand to the design and placing on the market of products and work equipment (responsibility of the manufacturer) and on the other hand to occupational health and safety (responsibility of the operator).

The European Machinery Directive [216], [217] is of great importance for products and work equipment. In Germany, this has been implemented at national level by the Produktsicherheitsgesetz (Product Safety Act – ProdSG) and the Maschinenverordnung (Machinery Ordinance – 9. ProdSV) based upon the ProdSG. The EU Machinery Directive is currently being amended in connection with the EU AI Act as the EU Machinery Products Regulation, which will result in extensive adjustments to standards. In the field of machine safety, harmonized standards mandated by the European Commission (see Chapter 1.4.4) are of particular relevance. The application of these standards triggers the presumption that the design of a machine complies with the legal requirements. Matters that are not regulated in harmonized standards must be evaluated as part of the risk assessment that is always required, and appropriate measures must be taken where necessary.

In addition to the Machinery Directive, there are a number of other European directives, including their national implementations, (or ones that are being developed, such as the AI Act) that must be taken into account in the technical design of AI systems.

In Germany, occupational health and safety is governed by the Arbeitsschutzgesetz (Occupational Health and Safety Act), which is essentially a national implementation of European occupational health and safety law. The central instrument of the Occupational Health and Safety Act is risk assessment, which focuses on the working conditions and the associated risks to the safety and health of employees. The German Occupational Health and Safety Act is substantiated nationally by ordinances, which are legally binding regulations. For further specification of the ordinances (e.g. the BetrSichV, ArbStättV, GefStoffV, ArbMedVV), government regulations are formulated in committees[86] established at the Federal Ministry of Labour and Social Affairs (BMAS) in an advisory capacity with the participation of the federal states, employers, trade unions, the German Social Accident Insurance (DGUV), science and, where appropriate, other institutions/associations.

In dual occupational safety and health, a coherent set of rules and regulations is being drawn up in Germany in coordination between the state and the DGUV, so that companies in Germany must also comply with the set of rules and regulations of the accident insurance institutions that are set up for specific sectors, or are supported by further detailed rules and information.

For the use of and work design of AI systems, central basic principles of prevention apply here in Germany via the set of rules and regulations for occupational health and safety. The basic principles of prevention are explained for Industrie 4.0 and to some extent for AI systems in the DGUV position paper 2/2017 [218]. Initial concretizations in the technical regulations of dual occupational safety and health in Germany are available and are being further developed on an ongoing basis. Projects or realizations of projects on the use of AI will require expert and specific reviews of the regulatory framework.

Aspects of data protection
AI systems typically use large amounts of data. This may give rise to interactions between data protection requirements and users' rights to privacy.

Basic principles of data protection such as

→ purpose determination or purpose limitation of data,
→ necessity,
→ transparency,
→ data avoidance and data economy

require careful consideration in the design of the sociotechnical system: Whether the principles are fulfilled in terms of content may depend on the technologies used, as well as the particular use case. Job application documents that have been discriminatorily presorted by the AI system for recruitment purposes (see above) could instead be used via AI to counter discrimination in the system.

Since data protection law does not yet provide a uniform set of rules for data collected in the employment relationship, company and service agreements that combine data protection-related aspects, technological approaches and value-based principles (e.g., code of ethics) can be useful in the company context.

4.4.2 Requirements and challenges

4.4.2.1 The sociotechnical perspective in the AI life cycle

The sociotechnical perspective must be considered throughout the complete AI life cycle (see Chapter 4.1.2.3). At each stage of the AI life cycle (cf. ISO/IEC 22989:2022 [16], ISO/IEC 23053:2022 [24]), the focus is on specific aspects of the sociotechnical system.

It is important to note that the approaches of general systems theory (in [219] & [220]) and sociological systems theory (e.g. [221], [222]) have only a limited effect in the development of AI systems. While a classical system does not evolve or evolves only slightly during its operational phase, AI systems have the ability to evolve within a set framework. Without careful consideration of this framework in the design phase, it can lead to unexpected and undesirable behaviour. During the operational phase, only “still images” of the system state can be captured, depicting the system and its sociotechnical interaction at a specific point in time. This complicates evalu-

[86] https://www.baua.de/DE/Aufgaben/Geschaeftsfuehrung-von-Ausschuessen/Geschaeftsfuehrung-von-Ausschuessen_node.html

ation and design, so special attention to system effectiveness from a sociotechnical perspective is needed.

In the following, the AI life cycle is highlighted with respect to the relevant sociotechnical issues in each case. Chapter 4.4.2.2 looks at the "Inception" phase, Chapter 4.4.2.3 at the activities in the "Design and Development" and "Verification and Validation" phases, and Chapter 4.4.2.4 at the "Deployment" and "Operation and Monitoring" phases, as well as the "Re-evaluate," "Continual Validation," and "Retirement" phases. It should be noted here that this does not claim to be a complete discussion of all sociotechnical perspectives and issues, as this would have gone beyond the scope of the Standardization Roadmap Artificial Intelligence.

When considering the sociotechnical perspective in the AI life cycle, it is also essential that the human and technical subsystems under consideration, as well as their interaction, are described and documented at each stage.

4.4.2.2 Initialization

This phase corresponds to the "inception" phase according to ISO/IEC 22989:2022 [16]. From a sociotechnical point of view, the goals of the application and requirements in particular are defined in this phase. For what purpose is an AI-based application needed? What requirements does the application have to meet due to the sociotechnical embedding? These are the essential questions that the relevant actors must ask themselves at the beginning of the AI life cycle and that do not require exclusively technical answers. Nevertheless, the answers specify according to which technical components should be selected in the next steps. Ultimately, it is a matter of moving from the idea to the decision for an AI system on the basis of an in-depth problem analysis in a given situation and initiating the development process. Consequently, this phase requires an intensive analysis of the AI system's field of action and its possible sociotechnical consequences. With the definition of goals and requirements, the framework for the further process is set and stakeholders can significantly align the development of the AI system at this point. In the next phases, there will be repeated references back to the parameters defined here.

Goal of the initialization phase

From a sociotechnical point of view, the goal of this phase is to contextualize the application in its environment. The first step is to make an initial decision as to

→ why an AI system should be developed,
→ which problem it solves,
→ which need of the target group is to be fulfilled, and
→ which are the success parameters.

Sociotechnical steps in the initialization phase

Define relevant groups of persons and involve them

It should be noted in particular that it is not what is technically possible that is decisive, but the real need that results from the problem analysis. Knowledge about the target group is essential and, with regard to diversity, should not reflect exclusively stereotypical ideas about people. The people affected and their rights are also building blocks for in-depth problem analysis. Involving these groups in the development process provides direct and unfiltered insights into their respective needs and can have a significant impact on the quality of the AI development process. As part of the BMAS-funded research project KIDD, an approach for the selection and participation of relevant stakeholders is being practically tested in various experimental spaces.

Participatory design approach

The involvement of relevant groups of people can also be achieved through a participatory design approach (cf. [229]). The main focus here is on jointly anticipating future scenarios. This can be condensed into the expression "reflection in action" (ibid.). The aim is to give users a voice without them having to become developers themselves (cf. ibid.).

Possible methods for this translation performance are the creation of prototypes, teaching models and simulations (cf. ibid.), excursions of similar, already running and functioning systems, scenarios, future workshops, games or "design fiction" [71]. This can be summarized by the term "storytelling methods" (ibid.).

Another important pillar of participatory design is the ongoing evaluation by users. Artificial intelligence applications pose new challenges to existing concepts of participatory design, particularly with respect to sustained evaluations of systems. This is due to the novel nature of AI components, such as the highly dynamic and interwoven nature of algorithms, parameters, and data, as well as statistical inference and the complexity of training datasets, which are not immediately apparent to those who were not involved in the design process.

In order to ensure safe use over the entire life cycle of the software, new participation concepts are required, especially for the ongoing evaluation. Initial approaches can be found in the work on XAI (explainable AI), which provides initial access to the underlying software logics, for example, via the disclosure of critical decision points or via new visualization concepts. These approaches require further elaboration.

Define sociotechnical requirements
In the initialization phase, moreover, requirements are defined overall that span the entire life cycle. Sociotechnical aspects in particular need to be defined in addition to technical requirements. The design requirements can also include ethical aspects. It is recommended that this requirement be operationalized in a way that results in these incentives for implementation. As an example, common catalogues of requirements for the development of AI systems from an ethical point of view can be ethics-by-design catalogues. Here, for example, we can mention the Algo.Rules developed by the Bertelsmann Stiftung, which use nine design principles and around 120 questions to define requirements along the entire AI life cycle [230]. In addition, the AI Ethics Impact Group's VCIO model provides an established method for operationalizing predefined ethical values [231]. Accordingly, values are broken down by criteria into clearly defined sub-aspects and these are made measurable by indicators and corresponding observables. In addition to the ethical aspects, the other sociotechnical requirements defined here in this key topic chapter must also be taken into account.

Analyse risks
Furthermore, this phase includes an initial risk analysis that identifies the sociotechnical consequences from the perspective of multiple stakeholders before the application is even developed and implemented (see ISO/IEC 23894:2022 [25]). In addition to technical and legal consequences, ethical and social consequences from the perspective of humans and society must be addressed accordingly. This results in a multi-layered significance of potential risks and risks to be identified. The following questions can help in this: "Which fundamental rights or values could potentially be affected by the use of the software? What are the intended impacts of the software? Who is affected by the use of the algorithmic assistance system? What are the potential impacts of using the software on different stakeholders? What are the potential impacts of its use on society, the economy, or the environment? Which risks could arise in the event of possible errors in the development or use of the software? What are the possible scenarios here?" [232].

Three criticality models
There are also essentially three variants of criticality classification for AI systems that can help sort out the multidimensional risks and recommend appropriate measures in a next step. Mokänder et al. [233] distinguish three common models:
→ the switch,
→ the ladder, and
→ the matrix.

The switch model acts as a binary classification approach. The AI Regulation proposed by the EU Commission uses the switch model by defining certain conditions to be met by a system that will later fall under the scope of the AI Regulation. This model is a relatively intuitive, low effort procedure, but it runs the risk of defining too few or too many systems for further referral [233].

The ladder model represents a higher level of complexity in this respect. This model distinguishes AI systems based on various factors and groups them into different risk classes. An already established ladder model is presented by the AI Ethics Impact Group – led by the Bertelsmann Foundation and VDE. The risk matrix presented there according to Krafft and Zweig (2019) distinguishes the intensity of potential harm from the AI system and the dependence of the affected person(s) on the decision in question (AI Ethics Impact Group 2020:35). Based on these factors, a distinction is made between five risk classes, each of which requires different risk management steps to be taken subsequently. The proposed AI regulation also establishes a similar risk matrix. The ladder models are united by the recognition that it is not the technical complexity but the modalities of social embedding that essentially define the risk of the systems. Although the ladder models are more complex, in practical application they open up sufficient guidance for classifying criticality.

The third model for classifying AI systems is the matrix model [233] and is a multi-dimensional approach. An example of this is the OECD[87] approach to classifying AI systems based on five dimensions. This model corresponds to the very diverse

[87] The Organisation for Economic Co-operation and Development [OECD(2022)]

use cases of AI systems and consequently presents the most complex model for classifying risks.

The three different models for classifying criticality each have their own advantages and disadvantages – especially in terms of practicability and informative value. In practice, mixed forms are conceivable; the AI Regulation, for example, works with both a binary approach and graduated risk classes. Depending on pre-defined conditions or dimensions, an AI system may or may not be defined as risky. The sociotechnical nature of the systems therefore requires a qualitative engagement with the models and a sensitive consideration of the potential risks to humans and society. Standardization approaches should take into account the multi-faceted nature of criticality.

Manage risks
The identified risks should be addressed in a next step with an appropriate plan. Here, already established risk management systems can support the reduction of the identified risks along the entire AI life cycle – in a process-oriented manner (see [25]).

Give detail to transparency and accountability requirements
In this phase, the demands for transparency and accountability are also constituted: What information must be disclosed? To whom must this information be disclosed? And with what technical depth must information be enriched in order to be helpful and understandable at the same time? Who can be held accountable for any damages that may occur? Without clarification of these aspects, further development of the AI system may lead to harmful consequences for humans and society.

Evaluate feasibility
In addition, costs, effort and resources are anticipated in this phase and the basic feasibility of the application is evaluated. ISO/IEC 22989:2022 [16] also defines commercial considerations in particular here. In addition, social trade-offs should also be considered in particular.

Once initialization is complete, further steps can be initiated within the planning and development phase. New information, for example about risks, may require a return to the initialization steps and should be incorporated into risk analysis and risk management, for example.

4.4.2.3 Planning & design

This chapter is concerned with the "Design and development" and "Verification and validation" phases of ISO/IEC 22989:2022 [16].

Goals of the phase from a sociotechnical point of view
In this phase, the AI system is given detail according to the previously defined goals and requirements (cf. Chapter 4.4.2.2). As a rule, several rough solutions are developed first, which are checked with regard to the fulfilment of the goals and requirements. Not every rough solution that is developed can fulfil the set goals optimally, so several planning loops may be necessary before the selected rough solution can be fine-tuned and all the necessary steps for commissioning and subsequent operation (see Chapter 4.4.2.4) can be prepared. It is important to involve all stakeholders (e.g., operators of the AI system, future users of the AI system, interest groups representing operators and users, representatives of civil society; for more details, see Chapter 4.4.2.3) in the planning process at an early stage and in a participatory manner (see, e.g., [203]).

In the planning and design of AI systems, the implementation of ergonomic fundamentals and principles and a usable design of products and work equipment are thus critical objectives for success. Thus, the application of these ergonomic fundamentals and principles is also an essential quality feature of AI systems as work equipment or objects of use. This applies throughout the entire product life cycle of the AI system (cf. Chapter 4.1.2.3 as well as Figure 19; according to ISO/IEC 22989:2022 [16]). From the product development process through commissioning and everyday operational use to decommissioning, not only the state of technological development and the specific use case must be taken into account, but also the fundamentals and principles of human-centred and participatory sociotechnical design. This requirement has not yet been reflected in the corresponding standards.

Contents of the phase
DETERMINE THE DIMENSIONS OF THE DESIGN
For the systematic and targeted design of the AI system, the underlying cause-effect relationships in the sociotechnical system under consideration must be known. The dimensions of design thus encompass all issues to be clarified in the planning and design process, delimit the legal framework (cf. Chapter 4.4.1.2) as well as valid standards and specifications, and indicate who is to be involved and which methods or instruments can be used.

When analyzing, evaluating and designing sociotechnical systems, it should be noted that they are always influenced by objective (technical-organizational) and at the same time human (personal) circumstances (e.g. DIN EN ISO 6385:2016 [235]). The HTO concept assumes that humans, technology and organizations must always be reflected in their interdependence and interaction. The work task plays a central role here, as it links the three elements of humans, technology and organizations [236], [205]. The dimensions of the design of an AI system therefore always result from the three elements human, technology and organization as well as from their interfaces (i.e. human-technology, human-organization and organization-technology) to each other.

Various action frameworks can be used to concretize specific issues. Table 8 outlines some relevant action frameworks with the design dimensions described there as examples. For more in-depth information, please refer to the respective sources.

Table 8: Exemplary action frameworks for giving detail to the dimensions of the design of an AI system

Author	Design dimensions to be considered
Bundesanstalt für Arbeitsschutz und Arbeitsmedizin (Federal Institute for Occupational Safety and Health) (publisher) [237]	→ Changeability (system, environment) → Transparency (experts, stakeholders) → Networking (internal, external) → Controllability (emergence, limitations) → Resilience (robustness, resilience) → Involvement of humans (actors, those endangered) → Consequences of damage (personal injury, other damage)
Huchler et al. (2020) [202]	→ Protection of the individual (health and safety, data protection and responsible performance recording, diversity sensitivity and non-discrimination) → Trustworthiness (quality of available data, transparency, explainability and consistency, responsibility, liability and system trust) → Meaningful division of labour (appropriateness, relief and support, agency and situation control, adaptivity, fault tolerance, and individualizability) → Conducive working conditions (spaces for action and enriched work, conduciveness to learning and experience, communication, cooperation and social inclusion)
IG Metall Executive Board (2019) [234]	→ Human-technology (adaptivity, transparency, complementarity) → Human-Organization (holistic nature, polyvalence, acceptance and participation) → Organization technology (decentralized control loops, optimization of interfaces)
The AI Methods, Capabilities and Criticality Grid [47]	→ AI component methods → AI component capabilities → Tiered taxonomy of a general risk assessment in relation to the system
ISO/IEC 12792 [238]: “Transparency taxonomy of AI systems” – project	→ Basic information → Organizational process → Usability of AI → Technical information → Quality and performance
The Fairness Handbook [224]	→ AI impact and risk assessment → Analysis of stakeholders and affected demographic groups → Fairness definition & metrics → Sociotechnical context study → Bias analysis

The identified design dimensions then result in various standards and specifications that are to be used for planning and design.

In the case of sociotechnical systems, dimensions of designing an AI system from an ergonomics/human factors perspective relate to

→ elements of the system (cf. Chapter 4.4.2.3) on the one hand, and
→ human-technology interactions (cf. Chapter 4.4.2.3) on the other hand.

These are examined in more detail in the following chapters.

Design requirements and recommendations from ergonomics/human factors currently relate primarily to static and stationary technical systems and equipment. On the one hand, existing requirements or design dimensions are not sufficiently described for new technologies (e.g. AI systems). On the other hand, additional requirements, for example due to systems that are dynamic in terms of content and time (such as AI, but also already simple mobile machines), have so far been insufficiently documented. Significant dimensions of this include:

→ digitalization (e.g. digital representations of real solution sets)
→ networking (e.g. variability of access width)
→ dynamization (e.g. variability in terms of time and content)
→ ambiguity (e.g. indeterminacy of the solution space)
→ degree of autonomy of the AI system (cf. Chapter 4.1.2.2).

Analyze the (sociotechnical) system in which AI is to be used

The elements of the sociotechnical system each represent effective execution conditions for a task processing by a human when using or deploying e.g. an AI system and can also be used as a taxonomy of different constellations of e.g. AI systems. Therefore, when planning and designing the AI system, it is necessary to conduct an analysis of the underlying sociotechnical system.

In the context of work, the sociotechnical system is the "work system" (cf. [199]). DIN EN ISO 6385:2016 [235] defines the work system as a "system comprising one or more workers and work equipment, acting together to perform the system function in the workspace, in the work environment, under the conditions imposed by the work tasks".

DIN EN ISO 6385:2016 [235] sets out principles of ergonomics in the form of basic guidelines for the design of work systems and defines the fundamental terms relevant to this. Ergonomics is the "scientific discipline concerned with the understanding of interactions among human and other elements of a system, and the profession that applies theory, principles, data and methods to design in order to optimize human well-being and overall system performance" [239]. Since ergonomics considers both human well-being and overall system performance, an ergonomics-centred analytical approach also includes the consideration of safety/security aspects, i.e., 1) "safety" (accident-relevant events) as well as 2) "security" (attack-relevant events). With regard to security, however, only an "inside-out" perspective is usually considered, i.e. threats emanating from the sociotechnical system itself (e.g. due to a lack of qualification of the people).

In a first step, DIN EN ISO 6385:2016 [235] provides for a requirements analysis to formulate the goals. Based on this, the following design fields can result according to DIN EN ISO 6385:2016 [235]:

→ Design of work jobs and tasks
→ Design of work organization
→ Design of work environment
→ Design of work equipment and interfaces
→ Design of workspace and workstation

The design fields outlined can often be transferred to other application contexts. However, this must be checked for the specific application and, if necessary, the system elements and design fields must be adapted accordingly.

General principles and concepts of ergonomics, which can be used for the design of sociotechnical systems, are specified in DIN EN ISO 26800:2011 [239], namely in particular:

→ Principles of ergonomics
  - Human-centred approach: adaptation of the components of a system to the characteristics of the user, taking into account the
    - target population
    - task orientation
    - environmental context
  - Criteria-based evaluation: evaluation of the application of ergonomic criteria
→ Concepts in ergonomics:
  - usability
  - accessibility
  - system concept
  - load-effects concept

- → Ergonomics-oriented design process over the entire life cycle

The design fields identified can then be used to derive the standards and specifications to be applied, as well as the relevant methods and instruments. In addition, the design fields largely determine which data is required for the design and subsequent operation phases and what the quality requirements are.

Furthermore, ethical aspects must always be considered in the planning and design of the sociotechnical system and must be designed for the entire life cycle of the AI system. Ethical aspects include transparency, accountability, privacy, justice, reliability, and sustainability (e.g., AI Ethics Label of the AI Ethics Impact Group [231]; see Chapter 4.4.2.2). In this phase, the sociotechnical system must also be analyzed with regard to ethical aspects in order to further specify the requirements identified in the initialization phase (cf. Chapter 4.4.2.2) and to adequately incorporate them into the planning and design.

The relevant standards specifications on AI (e.g. ISO/IEC 22989 [16], ISO/IEC 42001 [27], DIN SPEC 92001 series [162], [240], [117], ISO IEC 25059:2022 [35]), ergonomics and organization (e.g. DIN EN ISO 6385:2016 [235], DIN EN ISO 26800:2011 [239], DIN EN ISO 9241 series [514], DIN EN ISO 10075 series [513], DIN EN ISO 27500:2017 [271], VDI/VDE-MT 7100 [241]) and ethics (VDE SPEC 90012 [242], IEEE 7000 series [10], [11], [12], [13], ISO IEC/TR 24028 [28]) usually do not yet sufficiently consider the resulting requirements from the sociotechnical design point of view of an AI system and often disregard the interactions between humans, technology and organization. Therefore, these documents must be reviewed and supplemented as necessary.

Define division of tasks between humans and AI as well as the interaction process

The role of humans in the AI system varies depending on the AI technology used. Human tasks in the AI system as well as the resulting requirements and qualification needs can be derived, for example, from the three dimensions of the AI classification (for AI classification see Chapter 4.1.1.1), i.e.:

- → AI methods (classical artificial intelligence / knowledge representation and inference / machine learning / hybrid learning)
- → AI capabilities (perceive / process / act / communicate)
- → Criticality (no or low / certain / significant / substantial / unacceptable potential for harm)

Depending on the intended use of the AI system, there may also be other dimensions that affect the role of humans in the AI system.

Basically, with respect to human-technology interactions in sociotechnical systems, three hierarchically structured interfaces with respective design principles are of particular interest:

- → Task interface, e.g., as in DIN EN 614-2:2008 [181], the DIN EN ISO 11064:2011 series [243]
- → Interaction interface, e.g., as in DIN EN 894-1:2009 [244], DIN EN ISO 9241-11:2018 [245], DIN EN ISO 9241-110:2020 [246], ISO/IEC 29138-1:2018 [247]
- → Information interface, e.g., as in VDI/VDE 3850-1:2014 [248], ISO 9241-112:2017 [249]

Accordingly, a hierarchy of the design levels of human-technology interaction can be derived in accordance with Hacker ([250], [251]) (cited in Böde et al. (2013) [252]):

1. **Human-technology interaction (in the narrower sense):** Human-technology interaction in the narrower sense focuses on the design of task, interaction and information interfaces. The focus here is on the concepts of ergonomics (cf. [239]) as well as usability and user experience (cf. [246]). The DIN EN 614 [180], [181], [182], DIN 894 [515] and DIN EN ISO 9241 [514] series of standards specify the underlying principles and concepts and provide guidance for the design of human-technology interactions in the field of machines and systems, as well as consumer goods.
   The principles of task design are derived from the primacy of the task from ergonomics/human factors ([181], [253], [254]) and refer to completeness, scope for action, evaluability, variety, competence reference, contribution to results, development support, cooperation (cf. [255], [248]). The fundamentals of information presentation are explained in DIN EN ISO 9241-112:2017 [249], and these are discoverability, freedom from distraction, distinguishability, unambiguous interpretability, compactness, and (internal and external) consistency.
   The relevant interaction principles mentioned in DIN EN ISO 9241-110:2020 [246] are task appropriateness, self-descriptiveness, conformity to expectations, learnability, controllability, robustness against user errors, and user retention.
   It must be examined within the framework of standardization whether the existing standards already adequately reflect the new requirements of AI systems or if they need to be adapted accordingly. Standardization needs may

arise, for example, due to the dynamic allocation of functions and with regard to necessary strategies for averting the ironies of automation.

2. **Human-technology division of functions:** For the design of the division of functions between humans and AI systems, the primacy of the (work) task applies in principle, i.e. the design of the task is at the beginning of the design process and subordinates the design of the execution conditions to it ([195], [205], [181]). The procedure for designing work tasks is defined in DIN EN 614-2:2008 [181]. The chosen division of functions represents the degree of autonomy of the AI system (on degrees of automation, see e.g., [256] as well as Chapter 4.1.2.2).
   The relevant standards must be examined to determine whether they adequately take into account the various degrees of autonomy.
   The MABA-MABA list (= „men are better at" – „machines are better at"), which was originally developed by [257] (cf. e.g., [258], [175]), is sometimes used for the division of functions. In ergonomics/human factors research, this approach has been criticized as early as the early 1960s and has been discussed as an alternative ever since [259]. The manifestation of a fixed division of functions between the subsystems human and technology falls short, since it (1) postulates a mechanistic interaction of factors or subsystems, (2) generalizes skills, capabilities and knowledge of the subsystem human and does not consider their actual depth and interaction performance, (3) does not consider dynamics and further development of the subsystems, (4) does not consider the life cycle perspective for both subsystems, and (5) misses the objective of system design, whose success can at best be based on complementary supplementation [259], [260], [261]. A scientific appraisal of a complementary addition of the human and AI subsystems is still pending.
   In addition, the division of functions can adapt dynamically over the course of use depending on the situation (e.g., in decision-making or dangerous situations where the human must take over). This adaptivity is currently not yet represented in standardization (and is missing in general for automated systems, not only AI systems).
   In addition, in the context of dynamic function allocation, the „ironies of automation" can come into play (cf. [262]), showing that automation increases system complexity and therefore creates new tasks of monitoring, control, and correction for which human skills are often insufficient. This must be taken into account when designing the division of functions and automation and must be incorporated into the relevant standards and specifications.
   It is also necessary to examine how to deal with the concept of customizability in the context of AI systems (or automated systems in general). This is currently not reflected in standardization.
   Finally, the underlying guiding principle of technology design also plays an essential role in the design of the division of functions. If the human is seen as a source of error by the developer, the design will tend to try to reduce the influence of the human in the AI system to a large extent. If, on the other hand, the AI system is viewed as a support for humans, the division of functions will tend to be complementary. This should therefore be critically questioned at the beginning of the design process – a note on this should be included in standardization. In the interest of a human-centred use of AI, preference should be given to the guiding principle of a complementary division of functions.
3. **Prerequisites and consequences of human-technology interaction:** Finally, it is also important to design the organization and processes in which the AI system is embedded. A wide variety of aspects must be taken into account here, such as
   - trust in the system ("trust in automation")
   - stress and strain caused by the use of the AI system (e.g. technology stress), cf. DIN EN ISO 10075 series [513]
   - systemic effects (for example, cascade effects due to human intervention in the AI system)
   - changed risk compensation of the user as well as its consequences in case of (unnoticed) failure of the system
   - changes in the user's behaviour (e.g. with regard to communication or competence) and their consequences
   - the concept of culture in the underlying sociotechnical system
   - questions of responsibility and liability

In many of the aspects mentioned, the qualification of the users, participatory design and appropriate change management play a decisive role. It must be checked to what extent these aspects are mapped in the relevant standards and specifications (e.g., DIN EN ISO 27500:2017 [271], VDI/VDE-MT 7100 [241], DIN EN ISO 9001:2015 [263]).

Fine planning of the AI solution

After the framework conditions for the use of the AI solution have been clarified in the previous steps, all details for the use of the AI solution must now be finely planned, such as the selection of the technology and work equipment to be used, prospective risk assessment of the AI solution, creation of the necessary framework conditions in the company, qualification measures.

Preparation for operational use usually requires specific training of the AI solution. Here, special attention must be paid to the selection of training, validation and test data in order to avoid discrimination, etc. (cf. Chapter 4.4.2.2). Furthermore, the data to be used must be checked for quality with respect to the intended use (e.g.: enough data? inconsistent data? too old, too new data? wrong data?). In addition, selection of the datasets used and training, verification, validation, and testing of the AI solution must be adequately documented.

This phase must always be clarified specifically for the project in question and can therefore include a wide variety of aspects. As a rule, the specific product standards related to the technology used come into play here.

In addition, pertinent process standards may also be relevant, e.g. for the design of the organization [271], quality management systems (DIN EN ISO 9000 ff. [264], [263]), environmental management systems [265], energy management systems [266], and management systems for occupational health and safety [267]. It must be examined whether the relevant standards already sufficiently consider the use of AI solutions or need to be supplemented in this regard.

Plan deployment
After the fine planning of the AI solution has been completed, thedeployment must be planned, i.e. the initial operation of the AI solution. This typically includes scheduling as well as designing the communication process with stakeholders, and evaluation, feedback and mediation mechanisms for commissioning. Project management is usually decisive for this (e.g. DIN ISO 21500:2016 [268], DIN 69901 series [269], DIN 69909 series [270]. It should be examined whether AI projects have special features with regard to project management that should be mapped in standardization, if necessary.

In addition, process standards once again play a major role in this context.

Plan regular operation
Finally, regular operation must also be planned. This planning is based on the results of the AI fine planning, as well as on the planning of the commissioning, and thus also uses the tools, methods and processes defined there. Other planning aspects for regular operation can include, for example, aspects of operational personnel deployment, the establishment of a participation process for continual improvement, or the monitoring of the current development of the AI technology used with regard to change requirements. The specific planning requirements depend heavily on the particular application.

For the planning of regular operation, the relevant process standards form an important basis, e.g. for the design of the organization [271], quality management systems [264], environmental management systems [265], energy management systems [266], management systems for occupational health and safety [267]. These process standards usually do not yet sufficiently take into account the special requirements when using AI solutions and therefore need to be supplemented, especially with regard to the sociotechnical aspects.

Use of methods in design and planning
This Standardization Roadmap Artificial Intelligence does not claim to provide an overview of possible methods for planning and designing AI solutions, as these must always be selected on a case-by-case basis. During planning and design, it is therefore always necessary to check what the state of the art is for the specific planning case and to take this into account accordingly. The resulting standards and specifications are to be followed.

When planning and designing AI solutions, methods are required for, e.g. the:

→ process of designing AI systems
→ technical design of the AI system
→ ergonomic design of the AI system
→ design of interfaces
→ participation of stakeholders; supporting stakeholder deliberation on content through processes
→ technology assessment, damage analysis and risk assessment
→ evaluation, feedback and mediation
→ quality or result control of the AI solution
→ project management
→ communication
→ qualification, development of competencies, change management
→ documentation of the planning and design process

Stakeholders
Basically, the ideal typical requirement when planning and designing an AI solution would be to involve representation from all stakeholders. ISO/IEC 22989:2022 [16] divides such

stakeholders into “AI providers”, “AI producers”, “AI customers”, “AI partners”, “AI subjects” and “relevant authorities”.

Specifically, the following people may need to be involved, for example:

→ experts with domain knowledge (AI experts, data scientists, computer scientists, etc., process designers, usability experts, product designers, etc., software testers, ergonomists, psychologists, etc., security experts in the respective domain, experts in ethics, diversity, fairness, etc.),
→ in the company: experts from the departments concerned,
→ users of the AI system,
→ representatives of the interests of operators and users,
→ those making decisions regarding the deployment of the AI solution,
→ representatives of civil society,
→ and other perspectives.

The type, content and form of communication and participation depend on the respective point in time, related to the project life cycle, in particular

→ during target setting,
→ during planning and design,
→ during commissioning,
→ in ongoing operations or in the continual improvement process.

But it is not only the timing of the interaction that is relevant; it is also important to consider which stakeholders are involved in the interaction. Here, care must be taken to ensure that communication is always appropriate to the target group and is inclusive. Differences arise, for example, in communication

→ among experts with domain knowledge,
→ between experts with domain knowledge and users,
→ between users and technology,
→ between experts and other stakeholders.

Against this background, the processes of communication and participation must be planned accordingly and carried out methodically. Relevant standards and specifications (e.g., VDI-MT 7001:2021 [241]) can support this. It must be examined whether the relevant standards already sufficiently consider the use of AI solutions or need to be supplemented in this regard.

In addition, good practice examples or experimental spaces can provide support (e.g., the KIDD process (2022) [74]), moderated specification dialogues between experiential knowledge from the world of work, and technical and expert software knowledge [272].

4.4.2.4 Operation

This chapter addresses the “Deployment”, “operation and monitoring”, “continuous validation” and “re-evaluate” phases of ISO/IEC 22989:2022 [16].

Goals of the phase from a sociotechnical point of view
From a sociotechnical point of view, the goal in this phase is to ensure that the desired mode of operation defined in the initialization and planning and design phases is adhered to, as well as to decide at regular intervals whether the desired mode of operation needs to be adapted to changed framework conditions. For this purpose, real data (anonymized or pseudonymized, depending on the field of application) should be collected during operation and processed in a comprehensible and transparent manner for the relevant actors in the sociotechnical system. On this basis, a continual improvement process can take place, and people interacting with the system are provided with a resilient basis for making an informed decision about a possible intervention or other necessary measures. The stakeholders (e.g., employees in the corporate context) should be involved in the continual evaluation and adaptation of the system, and their experience with the system should be the basis and starting point for improvements [203].

This monitoring requires technical solutions that provide the necessary overview at all times in the sense of a transparency-by-design or transparency-by-default approach. This can be done either in a module within the AI system or through a stand-alone command tool.

Accompanying this, it is important to provide training on technical and AI-specific as well as interaction-based content (e.g., the effects of over-reliance or under-reliance) to those interacting with the AI system.

Actors involved in the operation of the sociotechnical system and their needs
Humans take on different roles during the operation of the sociotechnical system, e.g.:

→ the management of an organization in which such a system is in use;
→ persons in involved departments; works councils and other representatives of employees‘ rights;

- → users of the sociotechnical system;
- → IT specialists involved in the development and further development of the system;
- → interested parties from the affected or general public.

The roles that provide human oversight as defined by the High Level Expert Group (HLEG) are of particular importance:
- → the human-in-the-loop (HITL, involved in the AI decision cycle),
- → the human-on-the-loop (HOTL, involved in the design of the AI and in monitoring), and
- → the human in command (HIC, should be able to oversee the overall activity including broader economic, social, legal and ethical implications).

The role of the HIC is introduced in the proposed EU AI Regulation. HICs should have appropriate intervention capabilities, especially for high-risk systems, but basically for any AI regardless of its criticality, i.e. should be able to e.g. press a "stop button" for the AI [273]. The call for a "stop button" does not mean interrupting an ongoing AI procedure when doubts arise, but rather the possibility of not following a decision made by AI or suspending AI use for a certain period of time and letting humans decide instead.

Thus, human intervention should be envisioned when AI is used in sociotechnical systems. These could include, for example, allowing humans to make exceptions to the AI's decisions or reconfigure parameters of the system (thresholds, input variables). Both are conceivable as direct intervention by the users or, alternatively, intervention after involving authorized persons in the company. While the "keep the human in the loop" approach considers individuals in relation to AI, there is also the "keep the organization in the loop" design principle. This means that when using AI, the interaction of the relevant stakeholders should also be considered and continuously optimized [274].

In order to fulfil their respective roles or to satisfy their needs with regard to the transparency of an AI system, these actors require target group-oriented technical solutions – editions of a transparency or command tool, so to speak. These editions must differ in four respects:
- → In the depth and presentation of information: This will be different for the general public than for management or HICs.
- → In influence options: An HIC needs to have the ability to „intervene in the operation of the high-risk AI system or interrupt the system through a 'stop button' or a similar procedure," according to the draft AI Regulation for high-risk systems. For example, individuals in management may need to have the right to change target variables.
- → In the feedback options: In case of a (suspected) failure of the AI system.
- → And finally, in the qualification concept necessary to fill the role in the sociotechnical system and to use a transparency or command functionality to control and supervise a sociotechnical system.

The aspect of the organization in the operation of a sociotechnical system

The introduction of AI systems in the work process always means change for the introducing organization and its actors. The introduction of AI may require new skills, but it may also devalue existing skills. At the same time, tasks, roles and cooperation contexts can change. Against this background, it is necessary to identify resistance to change at an early stage and to address it. To make the change process successful, it is important to inform stakeholders and their representatives, to enable participation, and to create transparency and opportunities for influence. Accompanying organizational development is therefore already important in the initiation, planning and design of a sociotechnical system, so that it can develop the intended effect in the company. In the operational phases of transfer to the operational environment, operation and monitoring, continual validation and re-evaluation, accompanying organizational development also plays a decisive role.

Monitoring in operation: Transparency and intervention possibilities for humans in the sociotechnical system

The requirement for transparency for stakeholders, which arises from organizational development and also from the draft EU Regulation AI, comprises a number of building blocks: first, transparency about the defined goals and the intended mode of operation, i.e. the narrative behind the sociotechnical system [64]. Secondly, governance aspects, i.e. classification in the risk matrix as well as defined responsibilities, defined competencies and rights along the "long chain of responsibility" [63]. Presenting the results from the planning and design process, as well as explanations of the choice of specific calibrations of hyperparameters and evaluation criteria, in a transparent and understandable way is also important for monitoring in operation.

One example of this: Whether the results and functioning of a sociotechnical system in operation can be classified as fair depends on which aspects of fairness are relevant in the context. For example, in an AI system that suggests applicants for jobs, the distribution of different genders in the suggestions could be an aspect of fairness. So, according to the VCIO model, one value would be "fairness in terms of gender," and the associated criterion could be "percentage of men and women and other genders." One possible indicator would be that, based on the gender distribution among equally or similarly qualified applicants (e.g., 30 % women apply for this position), a proportion that is perceived as fair (e.g., 25 to 35 % women) also ends up on the proposal list.

In order to now perform monitoring when using an AI system, the decisions made in the earlier phases are important basics to have transparent:

→ Why were which goals chosen (e.g., fairness in terms of gender)?
→ Why were which target corridors chosen (e.g., 25 % to 35 % women on the job short list)?
→ What are the levels of autonomy, how are they defined and why? – so: At what point of deviation from the target corridor is more human intervention or even suspension of the AI system necessary?
→ Does a change to a lower or higher autonomy level take place automatically?
→ Which parts are defined as belonging to the AI system and must be able to be switched off completely in the case of high-risk systems, for example – and which parts of the system could remain in operation?
→ If a complete suspension of the AI system is enabled – i.e. a „stop button" is integrated: What might this even look like in the various use cases?
→ Can the sociotechnical system continue to perform its function when the stop button is pressed? If so, under what conditions and framework?

Standardization has the important task of defining a framework for clarifying these issues.

In addition to transparency about the defined goals, the design decisions and their backgrounds, what is now needed in operation is transparency about how the sociotechnical system actually performs in operation, i.e., systematic evaluation of the system's performance and risks in the field, as also required by the EU AI Act, Art. 61. Important issues are:

→ Which input data are used?
→ How do these change over time?
→ Is the quality still high enough in terms of the measures set in the planning and design?
→ Where are the specific measured values in relation to the target corridors, e.g., in the above example: What is the real percentage of a particular gender in the list of proposed applicants?
→ How does this change over time?

Depending on the target group and the depth of information, a simple traffic light system is helpful or more in-depth information is necessary. An intuitive and user-friendly interface design and a presentation that is understandable for the respective target group are critical for success.

The topic area XAI (explainable AI, explainability of the AI system) is of major importance in this context. The issues at stake are these: Based on which inputs are which outputs generated? Which aspects can be defined a priori, which can be identified with explainability metrics, and which can be collected ex post via target corridors?

Transparent information on targets and actual measured values are necessary prerequisites for monitoring a sociotechnical system in operation. There are other aspects to meet sufficient requirements as well: How is the dilemma of automation dealt with? Thus: Can humans even decide whether to switch off? Can humans still do the job without AI? What qualification is needed to enable them?

People's trust in human-machine interaction and the sociotechnical system as a whole is measurable. There are several ways to make such measurements. They differ significantly in terms of timing. The goal must be to identify when there is too little (under-reliance) or too much (over-reliance) human confidence in the automated process or decision, so that subsequent action can be taken: "Users must be able to make a clear mapping between the system functions presented through the interface and their goals." [275].

Training

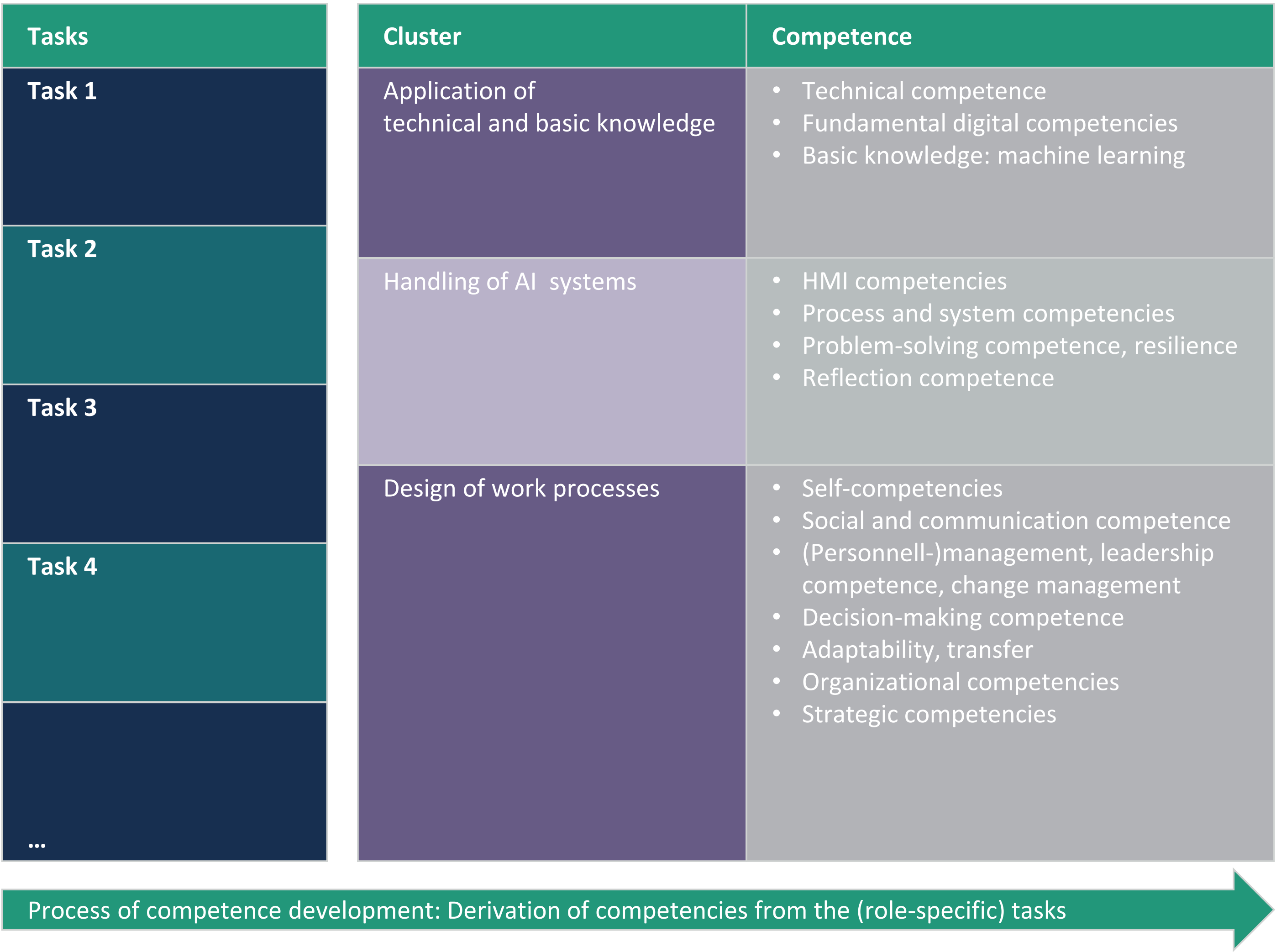
Different actors in the sociotechnical system have specific needs in terms of training: For one, training may be needed for the technical AI system. The way people interact with AI can also be an important issue: In this context, is over-reliance, under-reliance, or both to be expected? What organizational and social aspects should be included?

An overview of AI competencies and their development is provided in Figure 33.

Skills development is considered a key factor for successful implementation of an AI system. Training for employee competence development should be tailored to the employees' level of knowledge and the company's objectives. To achieve this, the first step is to determine which (job) roles arise in operations in the context of AI. These must then be formulated as tasks in order to derive the necessary AI competencies, see Figure 34. These competence profiles should then be assigned to the respective (job) roles. In order to take into account the employees' level of knowledge, an individual competence needs analysis should be carried out for their (job) role, based on the required competence profile, from which the individual need for further training measures can then be derived [190].

Iterative process: Continuous validation, re-evaluation and continual improvement

When planning a sociotechnical system, objectives and measures are also defined taking into account impact assessments. But not all decisions can be made a priori, as often not all necessary information is available, and framework conditions change over time. In addition, unintended effects could occur or it could turn out that planned measures were inadequate or incomplete. A change in the underlying data situation in the operational use of an AI system compared to the data situation at the time of AI system creation (training and test data) can be detected by methods in the "drift" aspect (concept drift, data drift ...) and should be a standard feature.

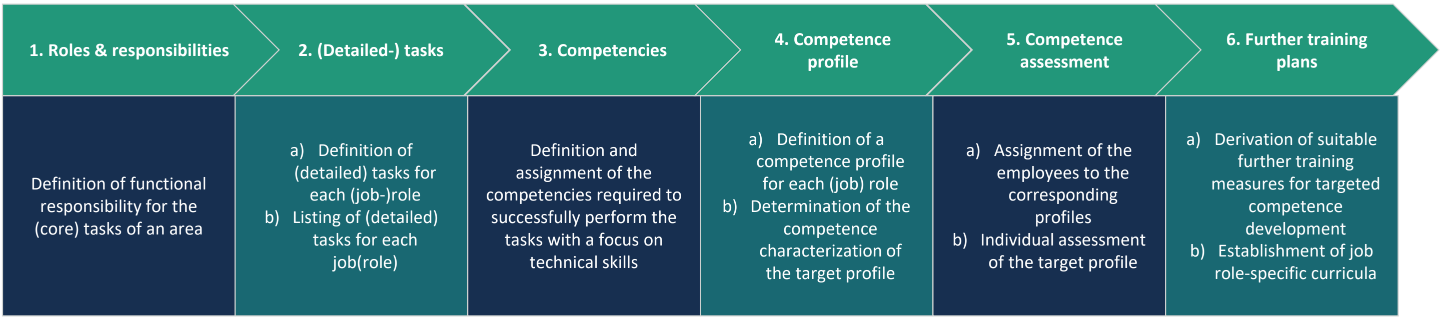


Tasks
Task 1
Task 2
Task 3
Task 4
...

Cluster	Competence
Application of technical and basic knowledge	• Technical competence • Fundamental digital competencies • Basic knowledge: machine learning
Handling of AI systems	• HMI competencies • Process and system competencies • Problem-solving competence, resilience • Reflection competence
Design of work processes	• Self-competencies • Social and communication competence • (Personnell-)management, leadership competence, change management • Decision-making competence • Adaptability, transfer • Organizational competencies • Strategic competencies

Process of competence development: Derivation of competencies from the (role-specific) tasks

Figure 33: Process of competence development and systematization of AI competencies (Source: based on [190]) Certification)



1. Roles & responsibilities	2. (Detailed-) tasks	3. Competencies	4. Competence profile	5. Competence assessment	6. Further training plans
Definition of functional responsibility for the (core) tasks of an area	a) Definition of (detailed) tasks for each (job-)role b) Listing of (detailed) tasks for each job(role)	Definition and assignment of the competencies required to successfully perform the tasks with a focus on technical skills	a) Definition of a competence profile for each (job) role b) Determination of the competence characterization of the target profile	a) Assignment of the employees to the corresponding profiles b) Individual assessment of the target profile	a) Derivation of suitable further training measures for targeted competence development b) Establishment of job role-specific curricula

Figure 34: Steps of a task-oriented competence management process (Source: based on [190])

For the reasons already stated, continual validation and evaluation of the goals and design decisions related to the AI system is necessary and even mandatory for high-risk systems under the AI Act, Art. 61. In the case of validation, the decisive questions must be answered anew in each case:

→ Are there additional/other goals to consider?
→ Do the existing targets and corridors still ensure proper functioning?
→ Or can there be a problem with the AI system even if targets/corridors are met?
→ If so, how can this be detected?
→ Do humans have to do this, can they do it at all, or what support is needed?
→ Is a corridor constantly being utilized in one direction and what „near misses" („near" failures) are occurring?

These "near misses" are often more common than true failures and provide valuable insight into the system when it has been operated at its limits. The introduction of reporting structures on failures or even "near" failures enables the analysis and improvement of AI systems to be able to ensure future system resilience or to simulate risks [276]. Again, this is necessary to implement in the proposed AI Act, Art. 62, at least for high-risk AI systems.

Since an AI is often used in interaction with other systems or other AIs, integration tests are necessary. In particular, such an integration test should be carried out before the first commissioning, but also during updates of the overall system. Here, the complete AI-based system is to be tested:

→ Are there side effects when the AI is integrated into the overall system (data formatting, timing)?
→ Are there any problems with the operation (usability) of the AI in the overall system?
→ Does the integration affect the performance of the overall system?
→ Does the integration affect the security of the overall system?

Feedback is another important basis for iterative verification of the sociotechnical system. These can be actively requested by the system ("programmatic feedback", rule-based), requested by an operator ("triggered feedback") [277] or reported in the form of a problem indication. The follow-up process (how and by whom is the feedback evaluated and what steps are derived from it?) must be clearly defined.

In the case of software updates as well as other changes of the AI system, a new alignment with the defined desired functionality is necessary. In addition, regression tests should be carried out with regard to performance, security and also usability. Any change can cause side effects that are not intended.

Ideally, it should be regularly checked whether there are fundamental changes in the AI technology that may have an effect on one's own solution or may achieve better results – or whether there are even other solutions instead of an AI in the meantime that solve the original problem faster or better.

The evaluation is followed by an optimization of the sociotechnical system. This can include individual components such as the AI used, the connections of AI with other IT systems and datasets, user interfaces, or qualification concepts. Characteristic of a sociotechnical approach is always the holistic view of the interfaces between the elements of technology, organization and humans and society. The continual improvement of sociotechnical systems targets these interfaces in particular. Examples of sociotechnical optimizations would be altered levels of autonomy or adapted human intervention capabilities.

4.4.3 Standardization needs

Need 04-01: Consideration of the dynamics of AI systems in the design of task, interaction, and information interfaces
In the planning and design of AI systems, the implementation of ergonomic principles and principles, as well as a usable design of products and work equipment are goals that affect success. Thus, the application of these ergonomic fundamentals and principles is also an essential quality feature of AI systems as work equipment or objects of use.

In the past, design concepts in ergonomics/human factors (including sociotechnical design) mainly referred to static technical systems (e.g. interface design to static and stationary machines). Not only, but also through AI (as a dynamic system in terms of content and time with cause-effect relationships that can no longer be documented), the EHF design concept must be extended so that dynamics of interfaces, modes of operation and effects are also designed to suit humans.

The relevant standards on ergonomics (e.g. DIN EN ISO 6385:2016 [235], DIN EN ISO 26800:2011 [239], DIN EN ISO 9241 series [514], DIN EN ISO 10075 series [513], DIN EN 614 series [180], [181], [182], DIN EN 894-1:2009 [244], DIN EN ISO 11064:2011 [243]) generally do not yet sufficiently take into account the resulting requirements from the sociotechnical design of an AI system and often disregard the interactions between humans, technology and organization in the interaction with AI systems. In addition, interaction concepts and information representation requirements are currently insufficiently mapped for self-dynamic systems for which continuous task processing is required and for which control interventions cannot be undone.

Need 04-02: Consideration of sociotechnical aspects in the design of AI systems
The way we work is changing with the introduction of AI applications, and so are the demands on humans. When introducing AI systems, organizational development, change management and the qualification of those involved are therefore important issues. In the sense of sociotechnical system design, the use of technology and the organization must therefore be planned and optimized together.

Relevant process standards, e.g. for the design of the organization (DIN EN ISO 27500:2017 [271]), quality management systems (DIN EN ISO 9000:2015 [264]), environmental management systems (DIN EN ISO 14001:2015 [265]), energy management systems (DIN EN ISO 50001:2018 [266]), management systems for occupational safety and health (DIN ISO 45001:2018 [267]) usually do not yet sufficiently take into account the special requirements when using AI solutions and therefore need to be supplemented, especially with regard to the sociotechnical aspects.

Need 04-03: Fulfilment of the standardization request for the EU AI Act, the aspect "transparency"
The draft **EU AI Regulation** (AI Act) places a focus on the sociotechnical perspective: The requirement to provide transparency and information for users can only be fulfilled if the AI system is understood as a sociotechnical system and humans are considered as part of the system.

Which transparency is sufficient in which context for which target group and which basic information must be available as a basis for human intervention in the system – these are questions that do not concern AI or AI developers per se, but rather the people who interact with AI.

In order to develop this standard, a broad involvement of the relevant stakeholders is crucial.

Need 04-04: Fulfilment of the standardization request for the EU AI Act, the aspect "human oversight"
The draft **EU AI Regulation** (AI Act) places a focus on the sociotechnical perspective: The requirement to provide human oversight can only be met if the AI system is understood as a sociotechnical system and humans are thought of as part of the system.

How human oversight is to be implemented in different roles and with a range of intervention options including a "stop button" triggered by humans, and what basic information must be available as a basis for human intervention in the system – these are questions that do not concern AI or AI developers per se, but rather the people who interact with AI.

In order to develop this standard, a broad involvement of the relevant stakeholders is crucial.

The Working Group Sociotechnical Systems ranked the identified needs according to the urgency of their implementation. Figure 35 shows the urgency of implementation according to the target group standardization.

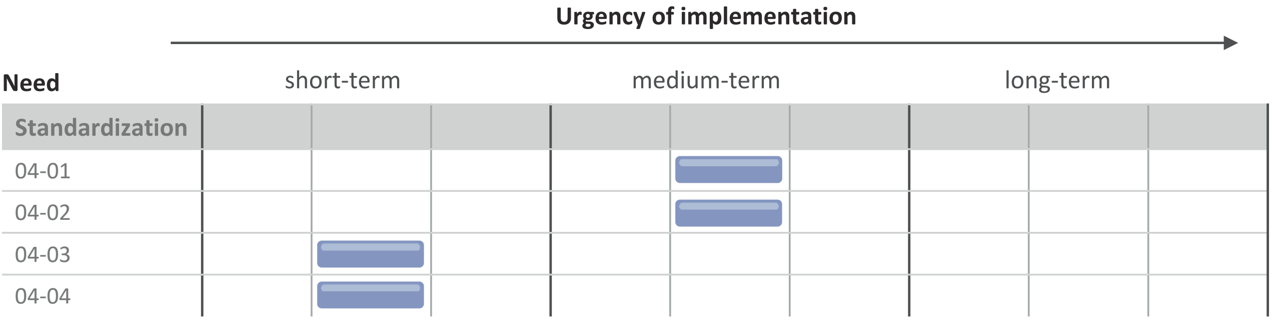


Figure 35: Prioritization of needs for the key topic sociotechnical systems (Source: Working Group Sociotechnical Systems)

4.5 Industrial automation

One-fifth of Germany's total gross value added is currently generated directly by the manufacturing industry and the processing sector; a large number of additional services are also dependent on this sector [278]. Accordingly, the manufacturing industry is a key driver of the German economy. As a result, the digital transformation of the manufacturing industry, i.e. the increasing use of methods and tools of information technology and (network) communication, is of essential importance for Germany as a centre for industry.

As part of the work of the project Industrie 4.0, which was initiated by the German federal government back in 2015, this topic area was worked on in a structured manner as a consistent further development of the automation of the manufacturing industry (Industrie 3.0). For this reason, the term industrial automation will be used synonymously with the manufacturing industry and Industrie 4.0 in the following. Corresponding future application scenarios were defined [281], [282], [283] based on existing value-added processes in the manufacturing industry [279], [280]. The application scenarios cover a broad range of applications, such as order-driven production based on dynamic value creation and supply networks, adaptive factories that enable the flexible adaptation of a factory's manufacturing resources, smart product development, and many more. These application scenarios form the basis for further refinements and analyses to derive possible research and standardization requirements [284], [285].

Artificial intelligence (AI) represents an important and essential key technology in the context of the digital transformation of manufacturing [285]. In particular, AI has a particularly high potential to sustainably design workflows and processes in the manufacturing industry [284], [287] and to increase value creation through dynamization and flexibilization, and to change business models in the manufacturing industry. Both traditional and newly designed production processes and secondary processes, such as logistics processes, can be improved, optimized and made more flexible through AI [282], [288]. In English-speaking and increasingly also in German-speaking countries, the term industrial artificial intelligence, or industrial AI, is used, which serves as a collective term for all fields of application of artificial intelligence in industrial applications [289], [290]. As the example in Figure 36 shows, different methods and algorithms of artificial intelligence can and will be used to implement different applications.

International standardization is of great importance in the manufacturing industry / industrial automation [285], [291]. In industrial automation, especially in the development and operation of automated systems, a large number of different manufacturers are involved; large supply trees for components and subsystems are common. Accordingly, interoperability across companies (mechanically, electrically, and in terms of software, communication and data, as well as their description) is of great importance, which is addressed by

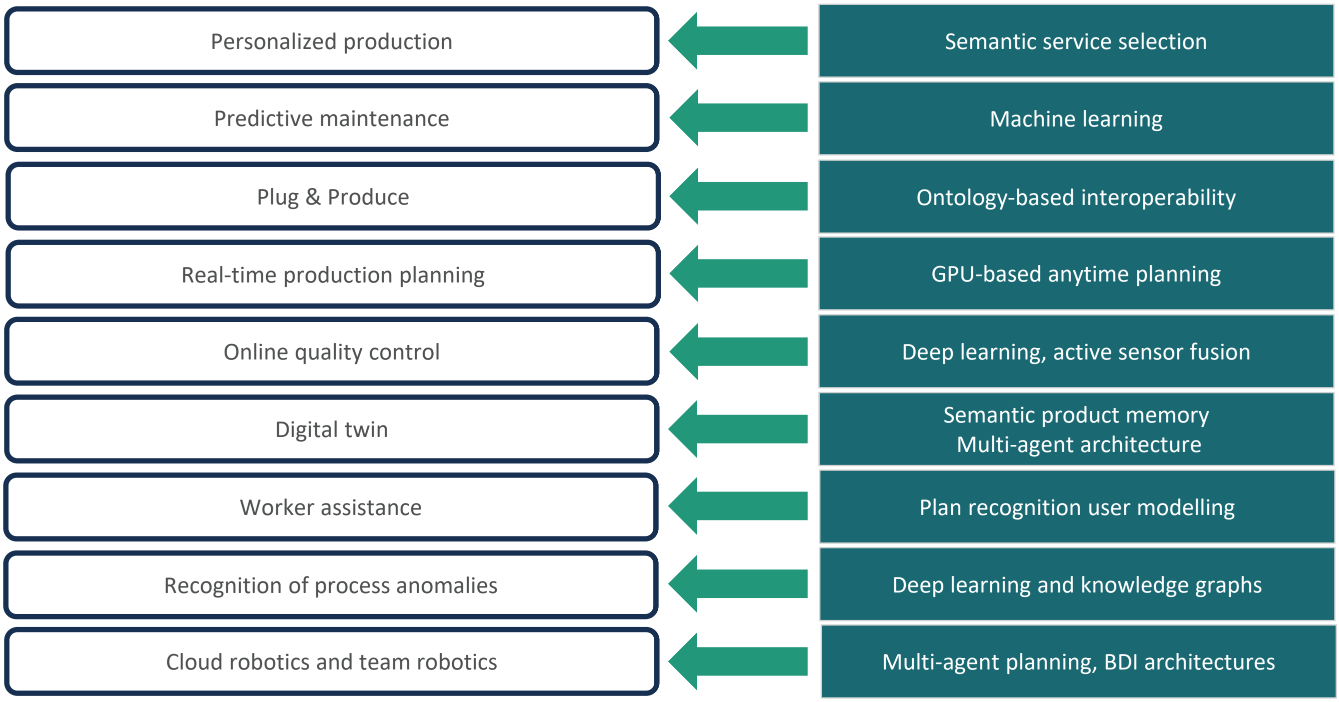

Personalized production
Predictive maintenance
Plug & Produce
Real-time production planning
Online quality control
Digital twin
Worker assistance
Recognition of process anomalies
Cloud robotics and team robotics
Semantic service selection
Machine learning
Ontology-based interoperability
GPU-based anytime planning
Deep learning, active sensor fusion
Semantic product memory
Multi-agent architecture
Plan recognition user modelling
Deep learning and knowledge graphs
Multi-agent planning, BDI architectures

Figure 36: Overview of AI methods and algorithms and their applications (Source: Prof. Wolfgang Wahlster, DFKI)

standards and specifications. Standards are also essential for defining solutions for compliance with regulatory framework conditions such as the Machinery Directive [216] or compliance with protection targets such as safe operation (see Chapter 4.2.1). Consequently, this also applies to the use of AI and thus underpins the standardization activities called for by the German government's AI Strategy [2]. For this reason, the topic of standardization for the manufacturing industry has already been examined in detail for many years in the DIN/DKE Standardization Roadmap Industrie 4.0, and dedicated needs have been derived; AI has also been explicitly addressed from the 4th edition of the Roadmap Industrie 4.0 [291] onwards and the standardization of AI in industrial applications has been continuously monitored [292].

In this context, according to the first edition of the Roadmap AI [63] there is a fundamental need for structured analysis of use cases and the derivation of normative requirements in industrial automation, which has already been addressed by IEC/TC 65/WG 23 and ISO/IEC/JTC 1/SC42/WG 4, with corresponding technical reports being updated (as in the case of ISO/IEC TR 24030:2021 [293]) or recently published such as PD IEC TR 63283-2 [294]. Consequently, the topic of use cases is no longer listed in detail in this current edition of the Roadmap AI.

An important role in the digital transformation is attributed to the digital mapping of physical reality: the “digital twin”. To ensure interoperability within a digital ecosystem, Plattform Industrie 4.0 is working with all participating institutions to develop the specification of the “administration shell” as a digital image of each relevant object (asset) in networked production [295], [296], [283]. An administration shell stores all the essential properties of an asset, such as physical properties (weight, size), process values, configuration parameters, states, and capabilities. In this context, the administration shell is not only an information store, but also a communication interface via which an asset is integrated into the networked, organized Industrie 4.0 production. This makes it possible to access and control all information in an asset. This provides the framework and an important foundation for the application of artificial intelligence for Industrie 4.0, as it allows data and metadata of relevant assets to be accessed in a uniform manner and to be made available in a structured data format. Current challenges related to data models and their semantics for the use of AI in industrial automation is considered in detail in Chapter 4.5.2. Well-known application examples for AI in Industrie 4.0 include predictive maintenance, where the service life and necessary maintenance time of components are predicted on the basis of symbolic as well as by means of machine learning models and collected operating data. To ensure the availability of necessary data (across companies), data spaces are becoming increasingly important in Industrie 4.0 [283] for the application of AI; here, too, explicit data models and their automatic processing (“reasoning”) play a key role.

The importance of explicit (semantic) models in industrial automation results, among other things, from their long-standing use in the development of machines and plants, which are often mechanically and electrically planned in detail and then automated by software. A large number of models are already emerging in this process, and their use shows great potential through the application of AI. Therefore, a not insignificant part of current activities in Industrie 4.0 is the development of technical systems in which artificial intelligence is used [297]. For this reason, the topic of “AI Systems Engineering” is considered and analyzed in detail for the first time in this edition of the Roadmap AI, and its relationship to standardization is described (see Chapter 4.5.1).

In addition, further applications of artificial intelligence are considered in the context of Industrie 4.0. In addition to autonomous intralogistics (see also Chapter 4.6), industrial image processing and image recognition, as well as the improvement of human-machine interaction and integration are also taken into account, for example. This is done through the use of new interaction mechanisms such as speech and gesture, new display capabilities such as augmented reality (AR), and the strengthening of collaboration through collaborative robotics, for example. Here, AI technologies find intensive application throughout.

4.5.1 AI systems engineering

4.5.1.1 Status quo

AI systems engineering addresses the systematic development and operation of AI-based solutions as part of systems that perform complex tasks [cf. Competence Centre AI Systems Engineering Karlsruhe CC-KING[88]]. Thus, AI systems engineering complements basic research on artificial intelligence (AI) and machine learning (ML) and bridges the gap to engineering sciences. The goal of industrial automation is

[88] https://www.ki-engineering.eu/en/what-is-ki-engineering.html

to make AI and ML methods usable according to the typical requirements and procedures of engineers, in safety-critical applications as well.

A major goal is the acceptability of AI methods, especially their use in subsystems of critical applications and complex systems. This requires a high degree of reliability, trustworthiness, safety/security and controllability. For this, there is a lack of accepted procedures and development methods in the industry. One first approach is PAISE(R)[89] – Process Model for AI Systems Engineering. Not all applications require the same level of these non-functional requirements. Therefore, a process model must be tailorable for a given situation. In addition, the question arises of how quality criteria can be described and validated in AI-based systems.

Since in practice AI methods have to be introduced into already existing systems (legacy systems), a process model should also support migration approaches and agile extensions of existing systems. In addition, AI systems engineering must consider the complete life cycle of AI-based systems, since deviations from the system context at the time of development can arise during operation, which must be systematically addressed. For example, the sensor data used by ML methods in operation may deviate from the training data used in the development environment in their statistical distribution to such an extent that the validity of the ML method is impaired (distributional shift).

AI systems engineering requires the coordinated interaction of representatives from different disciplines and backgrounds: engineers, AI experts and computer scientists. While the expertise and thus the technical requirements are typically covered by engineers (mechanical engineering, chemistry, process engineering, electrical engineering, ...), the knowledge of AI methods is mostly reserved for specialists (AI experts) who have dedicated mathematical and statistical methodological skills. Ultimately, an AI-based system is an IT system consisting of hardware and software components/subsystems that is to be developed and operated according to established system and software engineering methods. The necessary competencies are embodied in particular by IT experts with proven IT skills.

There is a visionary notion that AI systems engineering can deliver a method toolbox with clear statements about which capabilities, both functional and non-functional/qualitative, are achievable with which methods and under which conditions. For this purpose, it is necessary to be able to describe, evaluate and thus compare AI/ML methods according to uniform (meta-)models. This also includes the following tasks:

→ Development of a classification scheme for AI/ML methods: supervised vs. unsupervised ML methods vs. reinforcement learning methods; pre-trained ML methods vs. self-learning systems (reinforcement learning, Kalman filter etc.)
→ Description and validation of quality criteria in AI-based systems
→ Development of a domain-oriented explainability of AI-based decisions/decision proposals
→ Elaboration of the framework and meta-descriptions in structured and systematic data pre-processing, for structured, semi-structured and unstructured data, and for both static and dynamic data (including time series).

For most of these topics and tasks, rudimentary and singular approaches to solutions exist from science and industrial practice. AI systems engineering aims to bring these approaches and solutions together in a systematic and interdisciplinary manner and to establish cross-disciplinary standards that are accepted in practice for this purpose.

Use cases

The use cases described below are intended to describe and illustrate the interdisciplinary nature of AI systems engineering.

→ **Reliable energy supply for industrial production in the event of a fault**
For industrial production, it is essential to be able to rely on a reliable energy supply. Reliable energy supply means, in particular, avoiding cascading effects in the event of possible failure of distributed energy resources (DER) for energy transmission and distribution. The failure or breakdown of individual components can be withstood due to existing DER resilience and redundancy, while a cascading effect of failure of coupled DER components can result in a high risk of the blackout of entire network parts.
A cascade of failure of individual DER components, even under normal circumstances, can lead to a chain of further failures in power transmission or distribution. To

[89] https://www.ki-engineering.eu/content/dam/iosb/ki-engineering/downloads/PAISE(R)_Whitepaper_english.pdf

avoid the cascading effect, the following three factors must be kept under control by utilities [298], [299]:

- Attention to DER components because failures mostly happen in subcomponents.
- The power management system (PMS) must be able to remain fully in operation mode even under stress conditions.
- A possible cascading is an effect of the behaviour of very large systems („very large-scaled systems“ VLS, which are also called „systems-of-systems“).

It follows that the PMS must follow a new adapted “security policy”. To enforce such a policy, a measure is needed to classify technical stress, e.g. that caused by a thunderstorm. This categorization could be done, if necessary, by a suitable ML-based categorization procedure oriented, for example, to the severity of a thunderstorm. Furthermore, regulatory measures should be provided to increase the sensitivity of the PMS with respect to component failures with high risk for cascading. A practicable set of rules is based, among other things, on information and knowledge of how the failure of individual DER components influences the behaviour of the entire power supply network (failure propagation), e.g., in that blackouts of large supply areas or large industrial consumers can occur and thus cause great damage to the energy supply or production.

A blackout interrupts the power supply because the immediate compensation of the failure of a transmission line by other lines cannot take place due to indirect load overload or possible overheating, or the lines for compensation also fail.

A formalized scheme for the PMS requires computerized knowledge of high voltage transmission lines and transformers represented by nodes (graph vertices). Empirical values about transmission line failures, or “distribution factors”, characterize the possible consequences of a transmission line failure on other lines, which are plotted as graph edges on a graph.

Relationship to AI systems engineering: If AI methods are used in a subsystem of a PMS, e.g. ML methods for the optimization of energy networks in normal operation, early detection of anomalies in the energy network for the prediction of energy supply failures, as well as possible solution proposals for problem elimination or mitigation, statements on the reliability of such statements must be able to be made and justified on the technical level.

→ **Interaction of AI methods with the Industrie 4.0 concept of the administration shell**

The Asset Administration Shell (AAS) is an essential building block for increasing the value of a product. The complete life cycle of a product can be mapped and optimized by means of an AAS. The closer the modelling of a product corresponds to the „real“ object, the more accurately it can represent the „real“ object virtually and the closer one is to the idealized concept of a fully comprehensive digital twin. In conjunction with AI, process flows can thus be optimized virtually, for example, without having to intervene in real processes. For example, implementation in the physical world can only take place after a successful simulation in the virtual world using the digital twin.

A digital twin is understood here as a logical concept that fully maps the state and behaviour of a real asset in the virtual world. In practice, this is not economically feasible and, according to the use cases, not necessary. However, there is a vision that in a digital twin system, the data required for the respective use case can be procured via defined interfaces using a suitable service infrastructure of a data space, cf. the Reference System for Digital Twin Systems (DTS-RM).

Below are two examples of the use of AI in conjunction with the administration shell or digital twin.

Example 1: Use of AI in the digital twin for “predictive maintenance”

The massive expansion of the use of sensors (e.g. for temperatures, noises or vibrations) within a plant means that sources of error can be detected as they arise, thus minimizing downtimes. For detection, the digital twin – which reflects the target state – provides an optimal reference for the AI-based detection algorithm (anomaly detection).

Example 2: AI-based optimization of 5G campus networks based on a digital twin

The increasing demand for highly flexible production systems (e.g. batch size one) poses new challenges for planning and set-up. Moving parts such as automated guided vehicles (AGVs) must be fully integrated into the plant. The 5G campus networks offer a suitable solution. The use of a digital twin for a 5G campus network offers great potential for optimizing the entire system, both in planning and in ongoing operations. Appropriate AI methods are used for the optimization processes. By combining sensor technology and corresponding a-priori information, shadowing effects of 5G network coverage in a production hall caused by movement, for example, can be predicted and the path of the AGV adjusted accordingly.

Relationship to AI systems engineering: In order to reliably apply AI methods based on data from a digital twin, well-defined provenance and quality data of the digital twin is required. Since the data mostly originates from different sources, a standardized provision of the data and its processing is indispensable.

Standardization need
AI systems engineering is on its way to developing independently as a sub-discipline of systems engineering. This includes independent processes and methods that are scientifically based and generally accepted in practice, so that requirements for the necessary quality and the non-functional properties of a system can also be derived from them. These will be incorporated into specifications and also used in regulations. One example is the emerging AI Regulation of the European Union (EU) Commission, which will derive and make binding rules for the use of AI procedures based on the criticality of a system.

Generally accepted, coordinated and preferably standardized methods, models and approaches are required to implement these requirements and regulations efficiently and in a legally compliant manner. This is necessary because AI systems engineering can only succeed in an interplay of actors from engineering disciplines (mechanical engineering, electrical engineering, chemistry, process engineering, ...), computer science and data sciences [300]. The prevailing standards and specifications in these existing disciplines cannot be adopted unchanged, but must be conceptually merged. This includes a uniform definition of terms and the description of non-functional system properties that are also achievable and, if necessary, certifiable through the use of AI processes in subsystems.

4.5.1.2 Requirements and challenges

The following topics for standardization needs are derived from the described challenges of AI systems engineering:
- → Metadata descriptions of input/output datasets of ML methods
- → Metadescription of AI methods
- → Taxonomy, textual and, if necessary, formal description of quality criteria of AI-based systems, including reliability, dependability, plannability, controllability...
- → Metrics on explainability
  - Goal: Description of the trade-off between the explainability of the applied machine learning methods (thought model of the user) and accuracy or quality
  - Use of semantic models in explainability
- → Systematic approach to the use of AI methods in subsystems of complex systems (AI systems engineering approach model)
  - In all phases of the life cycle of a system (design → realization → operation and maintenance including enhancements/modifications)
- → Modelling of AI-based systems (technical and application aspects), e.g. Unified Modelling Language (UML) profile, special stereotypes for AI aspects

4.5.2 Data modelling and semantics

4.5.2.1 Status quo

A typical approach to AI data modelling today is based on the exploration of historical data. In this regard, initial methods such as the Cross Industry Standard Process for Data Mining, as proposed in AI systems engineering, have been widely established to perform the processing in an industrial context and achieve model building with the corresponding optimization. After implementation and the necessary testing of an AI system, the transition to operational use takes place, and downstream monitoring ensures that the data and data model, as well as the routines used are operated in the quality then required. Unresolved questions are, for example, how quality can be achieved with historical data that were created under different boundary conditions (Chapter 4.5.4, Need for action 05-09). By methodically extending systems engineering to include aspects of AI, for example, through the proposed AI systems engineering, industry can profitably use this new technology. Especially in modelling, the approach of putting elements and quantities into a context and explaining them in diagrams and figures has become established in industry. For example, well-known semantic data models are the entity-relationship model or the Unified Modelling Language used in object-oriented modelling. Abstraction for model building is necessary to represent the real world in the digital space (see Figure 37).

Conversely, the digital space reacts to the physical world, for example via “robot process automation” mechanisms. Data modelling involves considerable manual effort and is subject to a certain degree of arbitrariness. In addition, the challenge in modelling is that the system boundaries of the models are dynamic, and there is a great need for coordination in the creation of the models. This is ultimately reflected in the range of applications from highly specialized algorithms to general solutions. In doing so, modelling is fundamentally al-

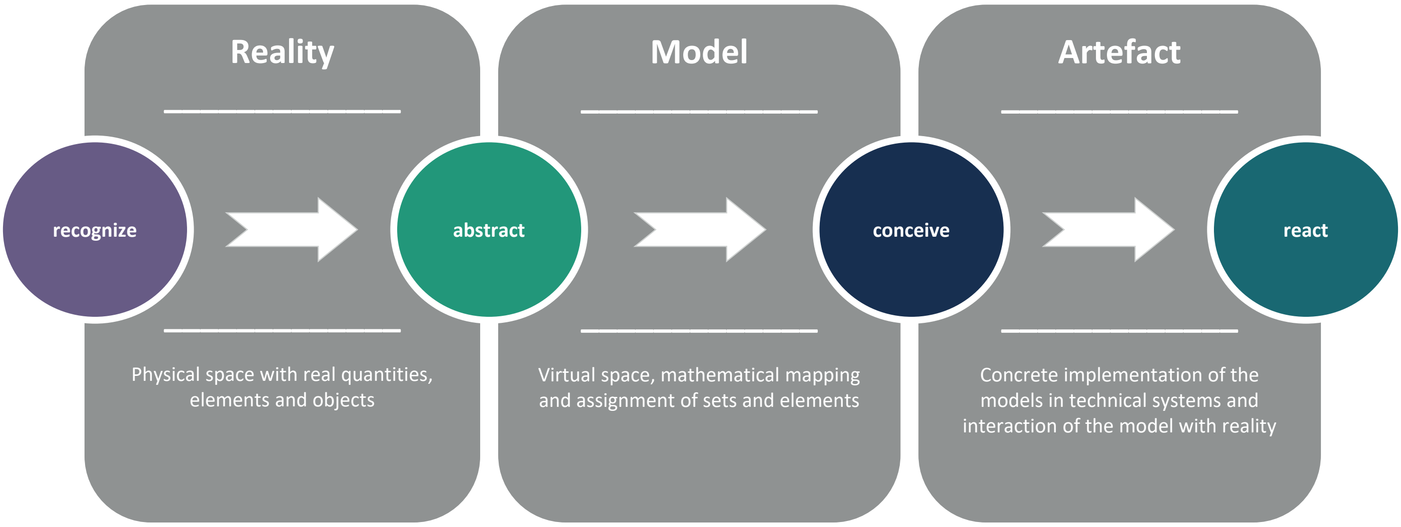

Reality
Model
Artefact
recognize
abstract
conceive
react
Physical space with real quantities, elements and objects
Virtual space, mathematical mapping and assignment of sets and elements
Concrete implementation of the models in technical systems and interaction of the model with reality

Figure 37: Model-building (Source: Working Group Industrial Automation)Systems)

ways about resolving contradictions between reality, models, and artefacts. What takes on importance are the interpretive mechanisms (cf. Figure 38) that operate from the physical space into the digital space and vice versa. The quality of both the data itself and the data models and their architectures is also underestimated, but this is necessary for successful implementation. Here, standardization can provide good support through metrics and standards for data quality (Chapter 4.5.4, Need for action 05-09). The power of machine learning lies in transforming data without the need for a complete mathematical prescription ahead of time. This is also known to be the disadvantage, which is specifically discussed in the explainability and predictability of the results today.

The importance of semantics, data modelling, data quality, and the interaction between the real and cyberphysical worlds is illustrated by the following general example of fire extinguishing. A first question is by which sensor technology "fire" should be detected and the context this use case has in the "real" environment. "Fire" in human interpretation could be sensed, for example, by a temperature sensor, a carbon monoxide sensor, or a camera. The generated data is completely different in each case, because it can be in degrees Celsius, particles or directly coloured or colourless images If only one data model may be generated, it must be able to handle the respective sensor data. The user is already faced with a decision here to determine the further course of action. Either each sensor receives its own data model or the different sensor data are interpreted and transferred into one data model. In practice, available sensors are replaced over time by newer sensors – a not uncommon scenario. But the data model itself could also adapt to the different sensor data. It is unclear today how verification will be ensured. Additionally, in order to provide an appropriate response to the event by the system, context must be established. This context is represented by the allocations. The fire can be the fire of a lighter or a burning vehicle – in that there will be different responses or reactions to it. Different actuators can also be used to implement the responses – in this respect, similar conditions apply as for sensor data. In the following, the example could be a burning vehicle.

In Figure 38, the "real world" consists of a vehicle in a certain dynamically changing state. The digital space (possibly a digital twin with an AI component) can simulate the dynamics in the real world in its process and data models and thus analyze them. The "vehicle on fire" incident must be detected quickly and correctly in the digital space and generate a fast and adequate response. Thus, an incident must lead to an appropriate response, because they are semantically related to each other as an implication. The cause-effect interpretation is "there is a fire, it should be extinguished". However, this interpretation can also be replaced by another interpretation "there is a fire --> heating in order" and correspondingly different context. The digital twin will be able to use its models to quickly analyze that the first interpretation is "correct" in a vehicle and the second may have its validity in a building.

In the digital world, the real objects, in this case the vehicle, could be made available through an administration shell and the usage could be integrated or monitored based on declared semantics, i.e., respecting axiomatic conditions. In this

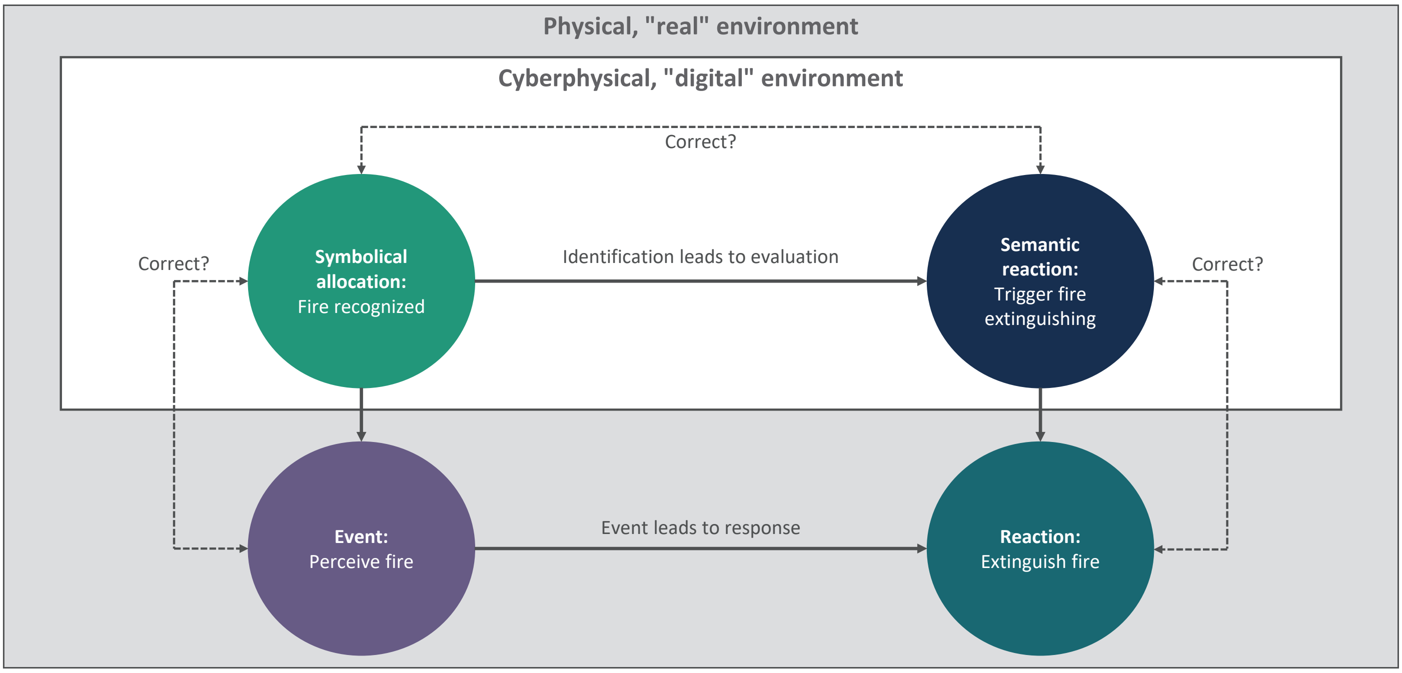

Physical, "real" environment
Cyberphysical, "digital" environment
Correct?
Correct?
Correct?
Symbolical allocation: Fire recognized
Identification leads to evaluation
Semantic reaction: Trigger fire extinguishing
Event: Perceive fire
Event leads to response
Reaction: Extinguish fire

Figure 38: Interaction (Source: Working Group Industrial Automation)

context, compliance with the declarations in the administration shell can also be realized and administered via submodels. In the example of the vehicle with an AI component, at least two submodels are then needed: one for detecting the incident and one for analyzing the "correct" response. While a data model would likely be appropriate for the incident sub-model, a process model based on graph theory would likely be more appropriate for the analysis of the potential response. It is noteworthy to observe that in a semantic view of the considered operations, the facts entered in the submodels of the administration shell are sufficient. The verification of correctness or the interpretation of these facts could be performed by an authorized digital twin with access to the submodels of the administration shell from the outside. This shows not only the importance, but also the complexity of the design of these systems.

Now, the first version of the Standardization Roadmap AI pointed out the importance of declaration and narration in semantic modelling with respect to the interoperability of such systems, especially in their dynamic interaction. The term "declaration" – in the sense of declarative knowledge representation – refers to the machine-verifiable representation of the structure and behaviour of interoperable systems, which is intended to be free of contradictions, using axioms, facts, and rules while at the same time dispensing with procedural parts. The term "narration", on the other hand, building on this and like a "program sequence", can mean two things:

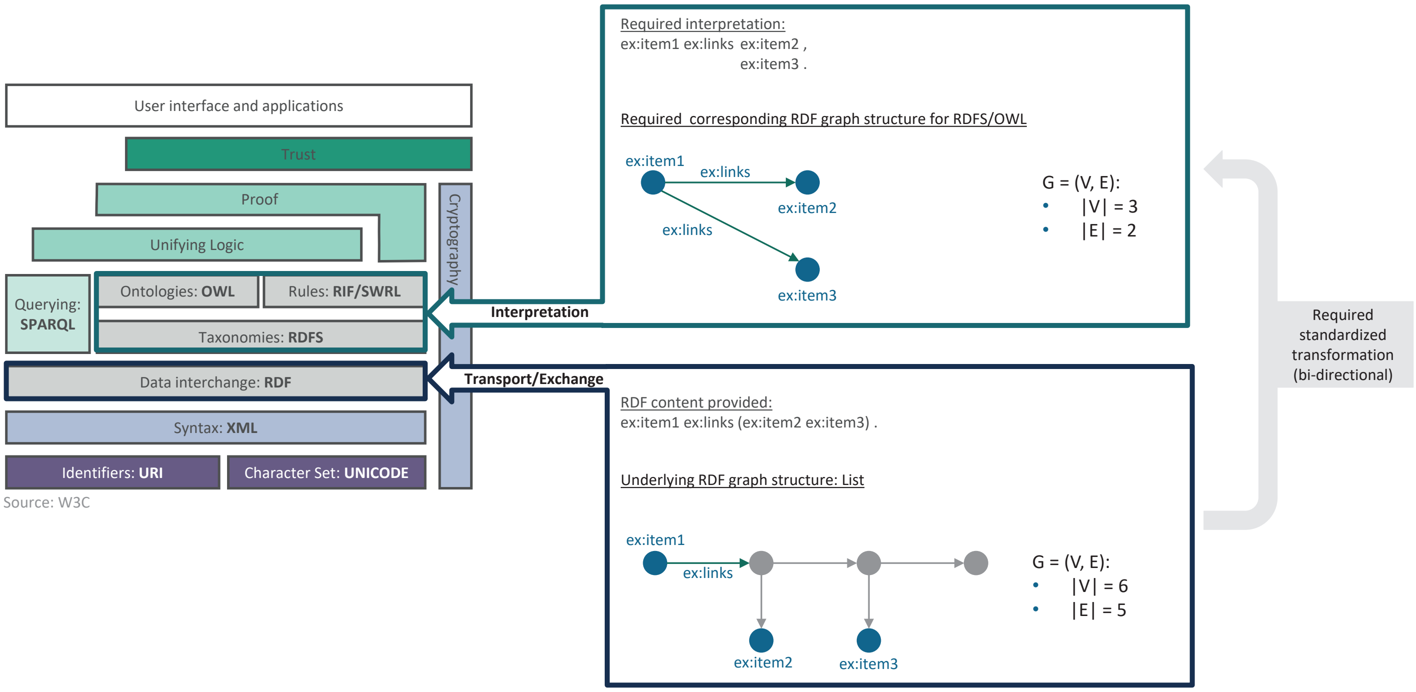
→ As a process, narration represents a dynamic orchestration of communicating, interoperable, model-based declaratively described systems with their variables. The respective orchestratability of the systems spans a search space of possible graphs of models through which a trajectory is formed by means of narration. In this respect, the concept of narration has a conceptual proximity to automated planning (see among others PDDL – Planning Domain Definition Language [301]): Plan Generation, Plan Execution).

→ As an artefact, narration represents a planned or, as a result of a narrative process, an actual sequence of interactions between communicating interoperable systems and their model-based defined variables or entities. Due to changes in the search space during narration, e.g., due to changes in the framework, especially at runtime, the concrete trajectory through the search space is formed. This can then be fed to further processing, e.g. validation.

In this context, the thematically related notion of "narrative" means narrative as an artefact and thus represents a declaratively described, reusable, verifiable trajectory through the graph search space. Thus, a narrative can be given a priori to influence a search space in a certain way ("What should be done?"). However, a narrative can also be determined a posteriori by observing changes to a graph ("What was done?"). In the case of interoperating (AI) systems, it must be assumed that the technical basis is different and not necessarily known to each other a priori, particularly in the case of dynamic interconnections, especially with regard to the coupling of heterogeneous systems. Moreover, different parties (systems, tools, knowledge engineers) can use the same models with a non-concordant interpretation of their semantics. This results in deviations and losses in the processing of the contents. Original intentions of data structures and models cannot be consistently expressed, passed on, and reconstructed. Thus, a lossless application of models and their validation for consistent interpretations across multiple parties along a processing chain ("pipeline") is not guaranteed (Chapter 4.5.4, Need for Action 05-06).

However, lossless, consistent horizontal as well as vertical interpretability of transmitted declarative knowledge is an essential requirement for the successful use of AI processes (Chapter 4.5.4, Need for Action 05-07). The transmission of knowledge takes place in two stages, once on the level of a suitable format and secondly on the semantic-interpretative level. A suitable format of knowledge is, among others, its representation as cause-effect implication, which can also be represented in programming languages. At the semantic level, these implications are represented as ordered pairs in a graph and are plotted as an event, i.e., a current edge, in the graph. Graph semantics are shared by the sender and receiver. Taking as an example the W3C Semantic Web Stack as basic building blocks that build upon each other: transported or delivered Resource Description Framework (RDF)-serialized structures cannot be fed to processing by RDFS- or Web Ontology Language (OWL)-based AI mechanisms in every case, namely if RDF serialization mechanisms are used that do not allow for standardized RDFS- or OWL-compliant interpretation. Even though RDF forms the basis for RDFS/OWL and the serialized content would in principle be RDFS/OWL compatible and serializable in its interpretation. The same is true when using the upcoming RDF* model. A concretely affected data structure is an RDF list, which can transport RDFS-/OWL-compatible unordered set elements, but for which there are no standardized transformations in the tools. This is shown in Figure 39.

In addition, tools perform individual transformations of the respective content when importing and exporting structures and models. These do not often preserve semantics, and are at the same time not verifiable, i.e. changed interpretations of contents cannot be recognized in every case. Transformation mechanisms of tools or systems cannot be addressed and tested in a dedicated way according to their capabilities. This means that it is not possible to recognize externally whether a tool or system can process offered content without loss (Chapter 4.5.4, Need for Action 05-08).



Source: W3C

Figure 39: Data modelling (Source: Working Group Industrial Automation)

The availability of corresponding transformations, which can be classified and also checked by third parties with regard to their semantic content, is therefore of central importance in order to make different (AI) systems interoperable. This is emphasized by [302]. The requirements of the FAIR principles (Findable, Accessible, Interoperable, Reusable) regarding interoperability [303] are necessary for this, but not sufficient. In addition, due to the associated higher complexity of model handling, reliable automation of the transformation functions is required. Design patterns can increasingly be used as a basis for this purpose. This becomes even more important with the emergence of data spaces as highly scaled "dynamic and automated meeting points" of (AI) systems against the backdrop of "trustworthy AI". This is also where a variety of interacting digital twins with individual capabilities and processes, typically provided by means of the administration shells, come together.

4.5.2.2 Requirements and challenges

The overarching challenge of "semantic interoperability" is found especially between heterogeneous technical or physical processes represented as complex, continuous time variables in the model. These variables and their characteristic properties are declared axiomatically and rule-based, for example, in the submodels of an administration shell. In interoperations between administered models, there is a flow of information of values between instances, objects or processes. This flow of values is thus the hallmark of interoperability between models or systems. When this stability is disrupted, the sequence of relationships is broken, i.e., the semantic facts between the models can no longer be satisfied. Interoperation should be possible between all models of different types with given semantics (Chapter 4.5.4, Need for Action 05-07) This means that an interface between models always has two sides, a sender side and a receiver side. Both sides use different technologies but share semantics, i.e. they operate in different "languages" or technologies but with a common understanding of, for example, information transfer.

This duality of an interface will be shown with a simple example, the interoperability between analogue and digital devices. Each device in itself has implemented the concept of "information" with its own means. While information in the analogue device is represented as voltage, measured in volts, as a function of time, in the digital device information is represented as a logical symbol "0" or "1" independent of time. In case of interoperation between the devices of different types, the current voltage level of the analogue device is converted into a symbol {"0", "1", "invalid"} depending on the analogue voltage level and the defined voltage ranges for {"0", "1", "invalid"} is transformed. This process of type matching (AD conversion) is called "coercion" and is performed automatically at the interface. As can be easily seen, there is a mapping rule for type matching in each direction with the two type parameters voltage level and voltage ranges. If all declared rules and axioms are observed, all forms can be continuously transformed into each other. The challenge here is to sufficiently declare all the required facts and transformation rules that must be adhered to or checked in order to maintain the flow of information. Based on the status quo, this leads on the one hand to the requirement for standards for declarative formats across models to establish semantic interoperability between models (Chapter 4.5.4, Need for Action 05-06). Furthermore, there is a need for standards for the representation of semantic characteristics of technical processes, especially involving AI-based components (Chapter 4.5.4, Needs for Action 05-17, 05-07, 05-08)

4.5.3 Humans and AI

4.5.3.1 General considerations

In Germany, the foundation of the legal framework is formed by the Grundgesetz (Basic Law) and made more precise in more specific laws and ordinances. There is currently no concrete design for AI systems and the EU has submitted a proposal for design with the draft Artificial Intelligence Act (AI Act) (cf. Chapter 1.4). Furthermore, there are already EU publications such as the 2019 "ETHICS GUIDELINES FOR TRUSTWORTHY AI", which also address the tension between fundamental rights and ethics in AI application. The guidelines state the three characteristics of lawful, ethical, and robust as the goal for trustworthy AI.

Technical standardization cannot answer questions about lawfulness and ethical principles, and the EU Ethics Guidelines also recognize that this is a political process of opinion-forming. In a standardization roadmap, it is therefore necessary to present the dividing line between technical standardization and the formation of political opinion in a recognizable way. In the context of technical standardization, this debate is less interesting because, with the exception of product safety, humans have experienced only indirect effects. In the case of product safety, however, the focus was clearly on the objective of not harming people, and technical standardization was able to make its contribution via the state of the art in product safety regulation. Standardization describes the process by which values can be incorporated and implemented in AI development and operation (cf. Chapter 4.1.2.2).

So when it comes to the robustness and security of AI systems, experts can get involved in the standardization bodies and develop procedures and requirements. For certain socio-technical systems and applications that intervene in society, issues beyond product safety and economics arise. Applications that initially appear to be unproblematic can, on closer inspection, have a strong social impact. One example is the automated processing of information and news in the context of forming social opinions. It therefore seems sensible that the dividing line for standardization work be determined and that the standardization bodies be given pointers as to the areas in which technical standardization cannot replace a democratic opinion-forming and legislative process.

First, it remains to be noted that an AI system (or algorithm) that merely demonstrates technical and economic improvements need not always have a direct impact on humans. Standardization does, however, help to create and maintain awareness of this if standardized process steps in the development and operation of AI reflect ethical implications (cf. Chapter 4.1.2.1). Therefore, the following list of questions, which can be used as a basis for an assessment, is aligned with the core question of how strong the influence and connection with humans is beyond the consideration of safety.

Is the AI or the algorithm being used to work together with humans to ...

1. manage and prepare information? Examples: News filters, censorship, statistics
2. manage or prepare templates for decision-making? Examples: Statistical evaluations, suggestion or optimization systems, health status assessment, scoring
3. make decisions? Examples: Knowledge bases, definition of medication, automation of administrative acts
4. execute decisions or measures? Examples: Automation of legal proceedings
5. exercise coercion or force? Examples: Automation of police measures; security and weapons systems
6. wage war? Examples: Security and weapons systems, strategic systems for warfare

The severity of impact increases (downward) in the list of questions, and there is another dimension to considering impact that is less obvious but immediately easy to persuade. This would be the question of the possibility of human cognition and influence on the algorithm used. In this context, there are differences in the roles involved in sociotechnical systems or systems that support the organization of society, such as those familiar from product safety. In addition to the familiar roles of operator, developer, instructor, and user, the system is extended to include the person concerned or the ordinary citizen, who may also be unaware that an algorithm has been active. From the distribution of roles, it can also be seen that there is a concentration of responsibility and power on the side of the roles that control the algorithm, and a lack of transparency and choice on the side of those affected by it. This is why this dimension needs to be addressed by further criteria on this imbalance.

The AI or algorithm works with data and metadata from and for ...

1. individual persons
2. small groups or communities of persons
3. large groups or communities of persons
4. at state or national level
5. transnationally or globally.

While the previous explanations refer to the circumstances and effects of the processing of data, which are taken into account in the EU Charter for Human Rights in Art. 8, the topic of data protection is also mentioned there, which in Germany is regulated in the General Data Protection Regulation (GDPR). Data protection is a broad field and cannot be comprehensively illuminated here. It is clear, however, that legally compliant data available for an algorithm does not simply exist and is subject to a wide variety of criteria, such as consent, limitation, transparency, oblivion and purpose limitation.

In addition, processing always raises the question of whether the data basis and algorithms used are free of discrimination and representative. And last but not least, whether the application of the findings on the individual[90] is relevant or

[90] For example, the AI makes a derivation from the data of all patients for the individual patient. This can (will) be maximally wrong for the individual, because they are an individual and not a statistical mean.

perhaps tangential to issues such as freedom of choice or presumption of innocence[91].

It is therefore recommended that the aspects shown above should serve as a basis for concretely defining the scope of work of technical standardization or for selecting the envisaged standardization projects in such a targeted manner as to avoid conflicts with the legal and societal aspects which are expressly not part of the work mandate of the technical standardization organizations.

4.5.3.2 Status quo

The application of AI in the human environment holds great potential and also raises many – currently unanswered – questions. By capturing, storing, and analyzing data, new abstracted contextual information can be automatically obtained and shared. Based on this digitization, AI goes one step further and enables processes and operations to act completely autonomously. The areas of application are very diverse and range from pure software applications and autonomous driving to medical technology, logistics or smart manufacturing. The realization of AI systems with increasing complexity is possible due to the constantly increasing computing power.

The topic of “Ethics/Responsible AI” was already discussed in detail in the first edition of the Standardization Roadmap AI. A key aspect is that people use artificial intelligence as a technology in a way that respects the rights and freedoms of natural persons. At first glance, this problem seems to be solvable, but a closer look reveals considerable areas of tension, since an AI is allowed to make independent decisions. In 1942 Isaac Asimov already (see also the AI narrative “from CYBORG to Digital Twin”) described basic rules of robot service in his short story “Runaround”.

Currently, the following problems can be identified:

→ A lack of transparency and explainability of autonomous systems
→ Questions about biases (prejudices) and the fairness of algorithms
→ Ethical AI design and the protection of privacy
→ Checking legal compliance, e.g. data protection regulation with monitoring tools, certification, etc.

To ensure the constructive use of AI as a technology, it must be ensured that it supports humans as a tool. That means: humans use AI in ways that preserve and enhance fundamental rights and freedoms. This also means: humans are not instrumentalized by AI. Here, there are concerns that AI will create a glass human being and that many jobs could be replaced by fully autonomous (AI-based) machines. In addition, questions of liability and the risk posed by the automated or even autonomous machine arise. For example, if an autonomous vehicle is considered, there is a relationship between the occupant, the manufacturer, and the AI. The vehicle is on the road on behalf of the human (the occupant), the AI has been developed by the manufacturer, and the behaviour of the vehicle is determined by the AI according to predefined rules.

From the perspective of anthropocentrism, the description of the capabilities of artificial intelligence in the literature is based on human-centred approaches, in which, among other things, human capabilities and senses are emphasized and imitated. Furthermore, there are approaches that dispense with anthropocentric approaches and describe artificial intelligence capabilities based on physical characteristics such as mechanical, electrical, and magnetic quantities.

Digitalization and digital transformation are among the reasons for the application of AI. Compared to digitalization, AI goes one step further and enables processes and workflows to interact largely autonomously according to the rules of semantic interoperability.

“Humans and AI” have a “duty to cooperate”, so to speak, in industrial applications in order to act as enablers. The duty to cooperate arises for good reasons, e.g., to avoid costs in the event of uncontrolled behaviour (see example of “fair play”), in the event of violations of process and data security, safety, etc. In addition to good reasons to use AI, there are of course also downsides. One example of this is the “deep fake” technology to generate imitations of people in images, video and sound. Distinguishing that with standardized criteria and metrics is one of the challenges in applying new AI technologies.

[91] Example “presumption of innocence” when an AI searches all citizens based on dragnet criteria and initiates measures without concrete evidence. Or simply to biometrically record all people at the airport in order to search for a criminal. This collection is a measure that happens without any reason for the individual.

A key weakness is the lack of a unified definition of AI and thus of an understanding of the different people involved in using it.

Although there is a standardized definition of artificial intelligence (see Chapter 1.5), it is formulated in very general terms. Consequently, it offers only a weak approach to concretizing definitions.

In the context of AI, standardization activities are already taking place at the international level, at the European level and at the national level (see Chapter 3.2).

Principles of ethical design of machines with AI components

IEEE P7000(™) ethically aligned design (EAD) principles
The ethical design of a software component such as a "digital twin" or the ethical design of a hardware component such as the "physical twin" within the framework of the three axes outlined "interoperations, plant hierarchies, value stream" of architecture models, such as RAMI4.0, SGAM (Smart Grid Architecture Model) etc., should be aligned with the recommendations of the IEEE P7000^{TM} series [64]. The P7000™ series contains so-called EAD principles for design and operational engineering activities such as performing tasks reliably (functional safety), critically evaluating failure analyses (functional reasoning), calculating behaviourally in advance (functional prediction), etc..

A similar role to that of the IEEE P7000™ series [10], [11], [12], [13], is played by the JTC1 SC42/WG3 standard ISO/IEC TR 24368:2022 [15] "AI Overview of ethical and societal concerns", which defines, for example, "fairness" in terms of behaviour or the assessment of outcomes.

The eight principles of an "ethical design" (P7000(™) EAD.ed2 [64], [104]), as developed by the IEEE, are limited to technical intelligent systems that are capable of autonomous behaviour. The capacity for autonomy should be able to place the interacting person in a state of "well-being and security" in many areas of life, stationary or mobile. This requirement is consistent with EU regulations under the AI Act, which specifically preclude avoidable or uncontrolled injury to the human "interactor."

EU regulations AI Act, DGA, DSA etc. in relation to "humans and AI"
At least since 2018, there has been a reflection in the EU on an EU strategy for AI [304], [305], which is described in a White Paper, published February 2020. The document "AI for Europe – An AI Strategy for Europe" [304] states that AI can not only make our lives easier, but it can also help address challenges such as treating chronic diseases, fighting climate change or anticipating cybersecurity threats.

AI therefore refers to "systems with intelligent behaviour" [304] as systems that analyze (i.e., try to "understand") their environment in order to achieve a specific goal. Once AI applications are working well and they have collected sufficiently qualified data, decisions can be automated, although with assessable risk.

At the DIN/DKE technical conference in November 2021 [305], the "EU AI Policy and the AI White Paper" and how companies can get involved were presented, e.g. with the question of how AI can be understood in regulatory terms by taking into account ethical citizenship and the **New Legislation Framework (NLF)**. To prevent subliminal tampering or avoid remote biometric identification, the AI Act calls for making an ex-ante conformity assessment for AI applications mandatory for all AI providers (suppliers).

In addition to the planned **AI Act**, the **Data Governance Act (DGA)** [306] deals with the availability of data and how trust in "data intermediaries" and their data-sharing mechanisms can be increased throughout the EU The data intermediaries correlate with the data spaces to be designed.

The goal of the DGA is to motivate stakeholders in all AI application fields to make their data available to the public for an otherwise worthwhile fee, while preserving their privacy rights and third-party rights.

The **Digital Services Act (DSA)** [307] similarly identifies a need for action arising from the regulatory framework for AI applications. It should be ensured that AI systems can be used safely and reliably in the market and that they respect, i.e. have implemented, the given fundamental rights and values of the EU.

All AI products and systems should be designed to give market participants confidence to make their investments and bring advanced AI innovation to market. Data governance measures and the enforcement of given laws, fundamental rights and security requirements for AI systems are to be adapted and improved. The European Single Market should be prepared for a common market for law-abiding, secure and trustworthy AI systems to avoid market fragmentation.

Technological-ethical narrative, from CYBORG to the digital twin
The hybrid organisms between humans and technology, "AI Machines (AIMs)," was invented in the 1960s when humans first attempted to leave Earth to assimilate distant worlds. "Cyborg and Space" may have been science fiction at the time, but it is now finding its renaissance in the use of AI technology in many industrial settings, the "smart spaces".

The Digital Twin, conceived as an **AI Machine,** was then as now a hybrid, not always transparent, possibly not trustworthy human-machine product that one wished to better contain.

Example: Thus, as early as 1942, Isaac Asimov, in order to prevent feared conflicts with AI machines (AIMs), drafted the first three ethical safety rules for hybrid AI components (AIMs) (or, the "three laws of robotics") in his story "Runaround" (which did not appear in German until 1982):

→ An AIM may not, and may not instruct itself, to injure a human being;
→ an AIM must always execute commands from the human, provided law 1 is not violated;
→ an AIM must be able to protect itself, provided that law 1 and law 2 are not violated.

The AIM example in this narrative serves to illustrate a need for action, recognized early in technological development, to maintain or not lose human control over complex non-transparent technologies at all costs. Today, there are promising approaches to answering these pressing questions with the various EU regulations, overlapping areas of digitalization, AI, ML, data use, distribution and security, etc.

With the use of AI-based robotic technologies, automation is moving out of the background of mere engineering and into the light of ethically-driven design. Automation has long since ceased to be just a task of regulatory technology for engineers, but is a social task as soon as automation opens up to questions of ethical regulation and the design of robots and mobile devices. New control concepts, such as the digital twin or the administration shell for valuable assets, e.g. industrial production facilities or industrial products, are being discussed in standardization and science today, or represent a need for action.

The task described elsewhere of controlling and optimizing the semantics or functionality of a RAMI4.0/SGAM-compliant system during runtime can be taken over by the digital twin [308], equipped with AI components, where necessary. The digital twin derives control signals from the actual-target comparison and sends them back as correction signals to the corresponding elements or artefacts in the (life cycle) domains and (hierarchy level) zones of the system under consideration.

4.5.3.3 Requirements for and challenges of the semantic AI-based systems

The semiotic triangle represents the holistic view of the semiotic triangular relationship between an anticipated thing, device, or asset, its ontological pre- or descriptive description of characteristics, and the most complete semantics or concepts of the thing under consideration. The relationship is called holistic because it contains the three representations necessary at least for understanding and their relationships to each other. It is the thing itself that obeys constructive requirements to function; it is the standards and lists of requirements to be able to manufacture things industrially with different machines in different places, and it is the understanding, in relation to language, that must be expressed to be able to build and dimension the semantics, functions, data flows, structures, including the use of the device.

Methodology of the semiotic triangle
The semiotic method for controlling automatic and autonomous processes is used, among other things, to represent de- and prescriptive requirements in ontologies and standards, properties such as trustworthiness, or the quality requirements for manufacturing products.

"Every product has three sides" that relate to each other and therefore must be presented or implemented together. The object or product under consideration (e.g., a heater that operates in a linear operating range of minus x to plus y degrees) consists of physical artefacts (components) and follows a physical function or serves a specific purpose (namely, it generates heat from electricity). This physical functional purpose is uniquely represented with semantic artefacts (e.g., nonlinear thermodynamic equations) and, possibly, solved for a particular workspace. In this solution task, the requirements of given standards and technical reports in terms of conformity, safety, reliability, explainability and traceability, etc. must be taken into account, especially in the case of nonlinear operations and processes or "non-safe" behaviour of AI components that engineers may use for the solution.

At the end of an "ethically aligned engineering" task for the production of a heater (to stay with the example) all three semiotic views are aligned: Clear (mathematical) semantics explain the purpose and function of the device, the standards provide a complete list of the (safety, quality) requirements to be met by the device and, if applicable, use cases for the operation of the heater. The device or type of device is designed by the engineer so that the device fulfils its purpose and functions reliably within its working range and throughout its life cycle [309].

Architecture of Choice

"Architecture of Choice" is the linguistic concept of arranging things, images, data, information that are available for selection in an automated way that makes it easier and possibly also more difficult for people to make a certain selection. For example, it is important for the person concerned which information and data are presented to them by an automated, semi-autonomous vehicle for decision-making purposes. This data must be able to be evaluated and decided by humans on a trust basis.

The opacity of built-in AI components and processes usually makes it difficult to confidently assess the results of an autonomous vehicle or automated production facility. Therefore, standardized procedures are required to estimate the risk of an intentional or unintentional shift of the target coordinates or the change of the result of a work or production step or the behaviour of a human at the human-machine interface (HMI).

Standardized procedures and metrics increase the comparability for users in determining the quality of products or services, as well as in evaluating a manufacturing process to have confidence in it or not.

At the HMI, people can be induced to make decisions that are in the interest of a hidden third party, e.g., a company behind the HMI used for cloud services, with intentional actions to influence their behaviour. Another example would be attempts to enforce societal interests of environmental or climate protection, etc.

One of the challenges of an "architecture of choice" or, in other words, an architecture to support decision-making at an HMI interface, is to present the choice options at the HMI interface in terms of the impact of the choices on present and future events around the people who have to make decisions. These people generally prefer "obvious" decisions that affect the current state to decisions that affect possible future states, such as avoiding an error state.

People or individuals differ significantly in their perception of information, and therefore draw very different conclusions from a comparable presentation of a selection of decisions to be made.

Another challenge of an "architecture of choice" is the open question of how human cognitive decision-making processes proceed and how architectural concepts can support and improve the quality of decision-making or run counter to it.

Smart technologies – smart capabilities of humans?

The development process recommended by standardization, based in part on EAD (see IEEE P7000(™) [64]), provides guidance on how values can influence the design of new technical standards through ethical consideration.

The term smart technologies is used above all in the field of the industrial Internet of Things (see JTC1 SC41 IIoT) to express that "things know their environment" because they are equipped with sensors and actuators, e.g. the "thing" of an automated production plant. The production plant can interact with its environment up to a certain quality. This makes it a smart production line because before it performs the next step, it checks all given safety requirements, e.g. no people in the safety area, in the HMI area.

Finished and future standards influence design, the implementation of processes and machines (of things) especially when using new technologies, and the benchmarks and transparency requirements required by EAD standards should be understood normatively.

The capability for semantic interoperability in complex **systems-of-systems** (very large-scaled systems) between things, machines, humans, models, and processes should be driven by standardization, science, and policy with standards, regulatory proposals, and research projects.

Maintaining a continuous exchange of data, values and knowledge, of information about energy or products, between cultures and systems requires control and governance of the **flow of added value** between different models, heterogeneous systems or cultures.

plannInformation models are mostly **layered models**, as represented e.g. in SGAM or RAMI4.0 and other reference models. Unlike semantic interoperability, which is about maintaining a flow of values, the layered model is about formats and protocols in more syntactic categories.

These HMI factors should be able to be used in heterogeneous system-to-system cooperation by applying human factor engineering (HFE), according to the new standard DIN IEC 63351, VDE 0491-61 [310].

4.5.4 Standardization needs

Need 05-01: Creation of a reference model for AI systems engineering
Creation of a common basic understanding of the terms as well as the interrelationships of the concepts used as an aid for the engineer in cooperation with computer scientists and data scientists.

Definition and explanation of terms and concepts and their interrelationships to systems engineering with special regard to the use of AI methods in subsystems; if necessary, construction of a formal model (e.g. UML, ontology, ...)

Need 05-02: List and definition of non-functional features (quality criteria) of AI-based systems, related to development and operation.
Creation of a uniform understanding for stakeholders (e.g. system requesters, system engineers), establishment of a uniform legal framework, creation of legal certainty for system behaviour and certifiability.

Definition and description of meaning for characterizing features such as acceptance, reliability, dependability, plannability, controllability, explainability, cybersecurity (security), functional safety (safety), uncertainty.

Need 05-03: Uniform approach to the evaluation of AI-based systems according to defined criteria
Definition of universal criteria and workflows for acceptance and comparison of the performance of AI-based systems.

Description of key steps in the workflow and application of assessment criteria, particularly for highly critical systems as outlined in the draft EU AI Act.

Need 05-04: Process model for the engineering and operation of AI-based systems
Development of guidance for the systems engineer on how AI-based systems should be fundamentally developed, operated, and maintained.

Definition of individual process steps for development, testing, acceptance, operation, maintenance. Description of the structure of the system and subsystems and AI-based parts. Details on the beneficial application of the agile vs. linear approach, defined design artefacts, details on documentation.

Need 05-05: Establishment of a standardized metadata description of AI methods
Creation of opportunities for building solution spaces (including catalogues) for requirement patterns.

Definition of a framework and classification of AI methods, formulation of semi-structured use cases, and derivation of potential AI methods to solve them.

Need 05-06: Characterization of data structures and models to use, preserve and reconstruct their original intentions
Different parties (tools, systems, knowledge engineers) should be able to use the same models with congruent interpretation of their semantics to avoid discrepancies and losses in processing. Thus, original intentions of data structures and models should be consistently expressed, passed on and reconstructed. This enables lossless application of models and their validation to consistent interpretations across multiple parties along a processing chain (“pipeline”).

Therefore, a validatable semantics of the intentions of structures and models across different parties along pipelines should be defined. To this end, robust procedures and mechanisms are to be described and defined by which the intentions of data structures, patterns, and models can be distinguished, preserved, and validated.

Need 05-07: Validatable transformations of structures and models
Transformation mechanisms of tools and systems for importing and exporting structures and models should be transparent and verifiable in order to be able to recognize changes in the transformed content as well as to avoid misinterpretations. This should allow tools and systems to be addressed and tested in a dedicated manner according to their capabilities. In addition, such behaviour offers the possibility to

detect from the outside whether a tool can process content offered to it without loss.

Transformation mechanisms should therefore validate their capabilities and data structures/formats by means of appropriate encapsulation, so that the semantics of the result can be seen in advance along a chain of transformations. For this purpose, characterization mechanisms and structures must be defined at the interface level, which can be used to understand, classify, and test transformation mechanisms.

Need 05-08: Identify and fix structural problems in the basic building blocks for compatible data/model exchange and AI
All processing levels along successive basic building blocks for data/model exchange and AI ("stacks") require holistic auditable conformity of the syntactic structures used. Currently, depending on the level of the semantics addressed, certain (data) structures are allowed or forbidden in stacks (example: the W3C Semantic Web Stack allows syntactic structures on the RDF level that are no longer permitted on the OWL level that builds on them). This leads to the fact that a "label of stack conformity" for tools and pipelines is not sufficient. However, content required to execute an AI mechanism must be contributed without losses and with auditable conformity of tools used with specified requirements and stacks.

For transformation mechanisms along stacks, test features are to be defined that can be used to automatically test content for usability or interpretability with the respective higher/lower levels of the stacks. For this purpose, it is proposed to examine stacks with respect to the vertically continuous interpretability of content and to define correspondingly standardized transformations for the respective bridging of stages. These should also be able to be combined with other stacks based on their semantics, so that semantics-preserving transport of content to AI mechanisms can be ensured across different stacks.

Need 05-09: Definition of metrics and methods for assessing data quality in ML data models, among others
Data quality is a decisive, and economic, influencing factor as soon as transactions are executed via the data models. Today, there is a lack of standardized methods to determine this characteristic, as well as metrics to evaluate data quality. A statement of data quality leads to a statement of model quality and thus to successful AI implementation.

It is proposed to introduce methods and metrics for data quality and to define mechanisms to validate this characteristic.

Need 05-10: Outline a specific I4.0 methodology for the design of I4.0 systems with AI components
The need for an I4.0 methodology arises from requirements for a uniform semantic view of I4.0 systems and of industrial plants, including data, processes and criteria for interoperability between humans and machines and machines-machines. This includes, among other things, linguistic expressiveness for the ontological characterization of a product or process.

The goal of the I4.0 methodology is to have a vocabulary with application rules, with which formal and computer-executable ontologies can be created and each "understood" (i.e., logically by the human and operationally by the machine) and used in their own particular way by humans and machines.

Need 05-11: Standardization and cataloguing of all artefacts categorized according to the thing-ontology/symbol-semantics scheme and their collection in stakeholder-specific catalogues for designers, developers, operators, etc.
The semantics of application scenarios should be able to be represented in a way that is both comprehensible by humans and executable by machines. This is the case with the use of graph and data types. Sequences of observable data-processing events are thus used to describe I4.0 manufacturing processes and products. Thus, the requirements of a particular narrative of an application scenario can be represented vividly as a graph trajectory on the one hand and semantically clear on the other hand. An example of an I4.0 narrative (formally represented as a graph trajectory with target state) is the "value flow" in the given reference architecture models.

An I4.0 methodology offers standardized tools and artefacts for designing application scenarios, use cases, or writing narratives, among other things. Narratives are characterized by linking a testable goal or validatable intent, such as successful quality control in production, to the manufacturing of a product. All steps taken to achieve the set goal must be documented. Metadata artefacts are available for this purpose. The comparability and usability of catalogued artefacts results from the set of rules applied (i.e., semantics) to design or create a thing or asset.

Need 05-12: Formalization of metrics, evaluations, testing, verification, and modelling
Since only rudimentary concepts of a common understanding or language can be observed in vertical standardization and in I4.0 industries, the evaluation criteria for testing functional safety or security requirements also drift apart. Therefore, there is a great need for action in the standardization of evaluation schemes and criteria.

A common language makes it possible to establish common standards of representation and evaluation. Common logic/semantics includes linguistic, ontological, and logical categories of artefacts, which can be used to verify, for example, digital-twin modelling or a semantically faithful, i.e., behaviourally correct, implementation (relative to the model) of cyberphysical reality.

Need 05-13: Test and evaluation methods for assets with built-in AI components to estimate the impact of AI on system or component quality
AI or ML are seen as tools that, when built into assets, can change the quality of the assets. This results in the need to examine to what extent quality changes have an influence on the functional safety of an asset.

Needs arise from assessing the impact of new (AI) tools and components, built into manufacturing equipment and products, in terms of human-machine relationships, e.g., on tasks to be performed together, on the quality of the products so produced, etc.

Need 05-14: Obtain arguments and metadata that can be used to substantiate the trustworthiness of the actions of stakeholders involved
Trustworthiness is not always only a problem of product quality measurement, but often a problem of product usage, in which transparency and self-explanatory input/output behaviour play a role. Therefore, there is a need for methods to verify the effectiveness of control over the product or production site.

The need to prove the trustworthiness of a product or process changes, among other things, with changing technologies and processes, e.g. when applied in a production plant. It is therefore a continual process of renewal, which also requires a continual review of the assurance of trustworthiness.

Need 05-15: Creation and continuous updating of a (semantic) standardization map with built-in help for using the map
Standards are often considered and written in isolation for the design of equipment and products, without deeper knowledge or references to other relevant horizontal and vertical standards, due to the lack of a semantic coordinate system for standards in the standardization landscape.

A standardized form of a common representation of semantics can be helpful in attempts to locate and, if necessary, check interrelated standardization topics in a heterogeneous standardization landscape, such as that resulting from RAMI4.0.

Need 05-16: Internationalization and digitization of standards for new technologies to support automated evaluation of system requirements
A standard written in “standardization English” is usually “translated” or “back-translated” several times by different stakeholders until it is complied with. A digital standard, on the other hand, can provide the requirements to be implemented in a way that can be processed in machines in a computational logic language. These parts may include data, processes, and “knowledge” databases for machine learning or decision-making.

Need 05-17: Development of a common (I4.0) language that allows a system to be described in different viewpoints but in a uniform rule-oriented representation
Typically, multiple logics (called viewpoints) are used simultaneously when writing standards and specifications. All logics have in common laws, axioms, rules of derivation for their specific domains they represent. Examples of “logics” are product liability, safety, security, privacy, functionality, interoperability, quality specifications, production dimensions, etc.

Need 05-18: Cataloguing of technical, semantic, and legal terms or artefacts for constructive synthesis of AAS submodels
It is necessary to define a common language (i.e., common representation of semantics) with rules, laws, axioms in an industry-specific manner, and parts of them in a cross-industry manner, in order to regulate and ensure, among other things, user safety in the cooperation between humans and machines controlled by an embedded AI.

Need 05-19: Standardization of the aspects of the "Human & AI" ecosystem
Humans & AI are evolving into an ecosystem with implications for social, economic, private, and workplace actions of humans and their cooperation with machines that rely on AI.

In order to make the ecosystem transparent, the following need for standardization arises:

→ Pointing out the mutual effects of AI vs. humans
→ Describe and define responsibilities of "AI" and humans in different roles and collaborations
→ Describe and define scenarios in the multidimensional interaction of AIs and humans

Need 05-20: Realization and implementation of the Digital Service Act (DSA) in the "Human & AI" ecosystem in various vertical applications and data spaces
The DSA contains 35 articles, grouped into eight chapters, giving guidance to build data spaces from private data ownership. In these data spaces, AI components can be used to analyze the available data.

The need for standardization arises for the data spaces required by industry, filled with economic, production and labour data from private-sector sources. This collaborative process between giving stakeholders and using stakeholders must be shaped with standards, regulations and legislation for the benefit of all.

Need 05-21: Standardization of origin evaluation methods and mastery of vertical data spaces
The narrative of AI shows that formalized rules are the basis for mastering and understanding complex processes, with the elements of AI-supported system components and data spaces made visible and explainable by means of metadata and meta-rules. Metadata and meta-rules represent knowledge about the creation of things and production data or about functions of things and use of data.

Need 05-22: New standardization projects for "formal and semi-formal" standards for the semantic concretization of technical topics and the behaviour of systems to be performed within the framework of technical standardization
Formal and semi-formal standards are standardization texts that are partially or completely "computerized", i.e. that can be processed and edited with a computer. One example of this would be the "digital twin".

Many topics and statements in standardization and regulation concern overlapping areas of competence. For example, the political opinion-making and legislative process is decisive for the regulatory design, which provides the concrete framework for certain technical use cases and objectives (e.g., devices that may also have a military use).

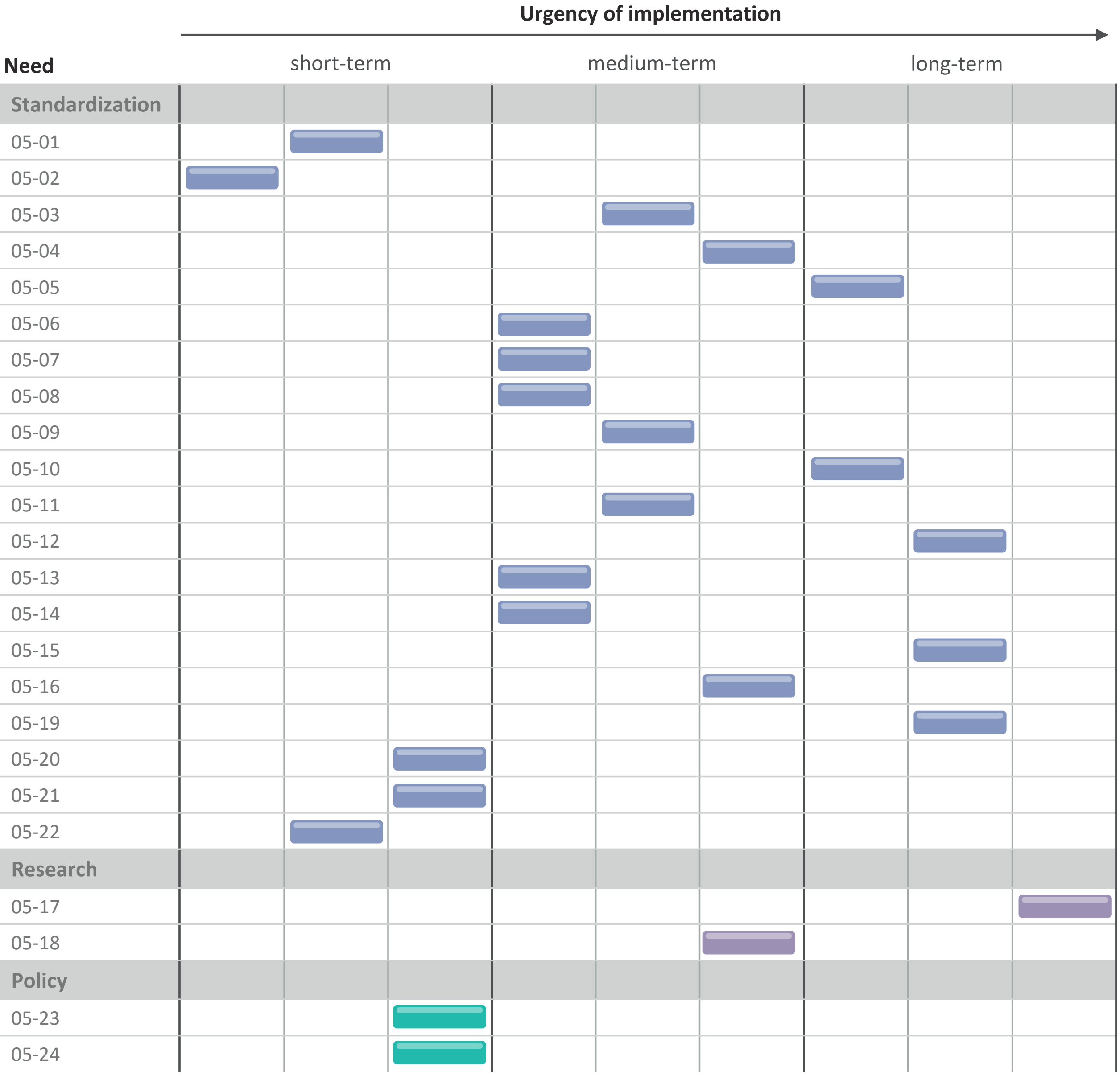
Need 05-23: Address the application of ethical rules as far as possible and in cooperation with democratic institutions nationally and in the EU in standardization
The issue of ethics and its application is part of the political opinion-making and legislative process. Ethical issues can only be partly or not at all worked out via standardization procedures. As a rule, they form a core aspect of democratic processes and are handled differently internationally.

Therefore, joint standardization efforts should be sought to find ethically-based (normative) procedures on regulations of AI biases, sludging, nudging, ethically-aligned design, etc.

Need 05-24: Clear definition of high-risk AI systems and their differentiation from safety systems
High-risk AI systems (in the sense of the EU Commission's proposed AI Act) can also be systems that are not considered safety systems. However, similar requirements will apply if it is desired by the legislature to design all high-risk AI systems as safety systems (in the sense of fail-safe, functional safety) in the future. AI and data model regulations complement the technical requirements specified in standardization, in the sense of a "red line" that should not be crossed ethically or legally.

In the context of the ongoing discussion about the planned AI Act, it must be clarified whether all high-risk AI systems will have to be designed as safety systems in the future, since the proposed legislation provides for different requirements for high-risk and safety systems.

The Working Group Industrial Automation has ranked the identified needs according to the urgency of their implementation. Figure 40 shows the urgency of implementation, categorized according to the target groups of standardization, research and policy.

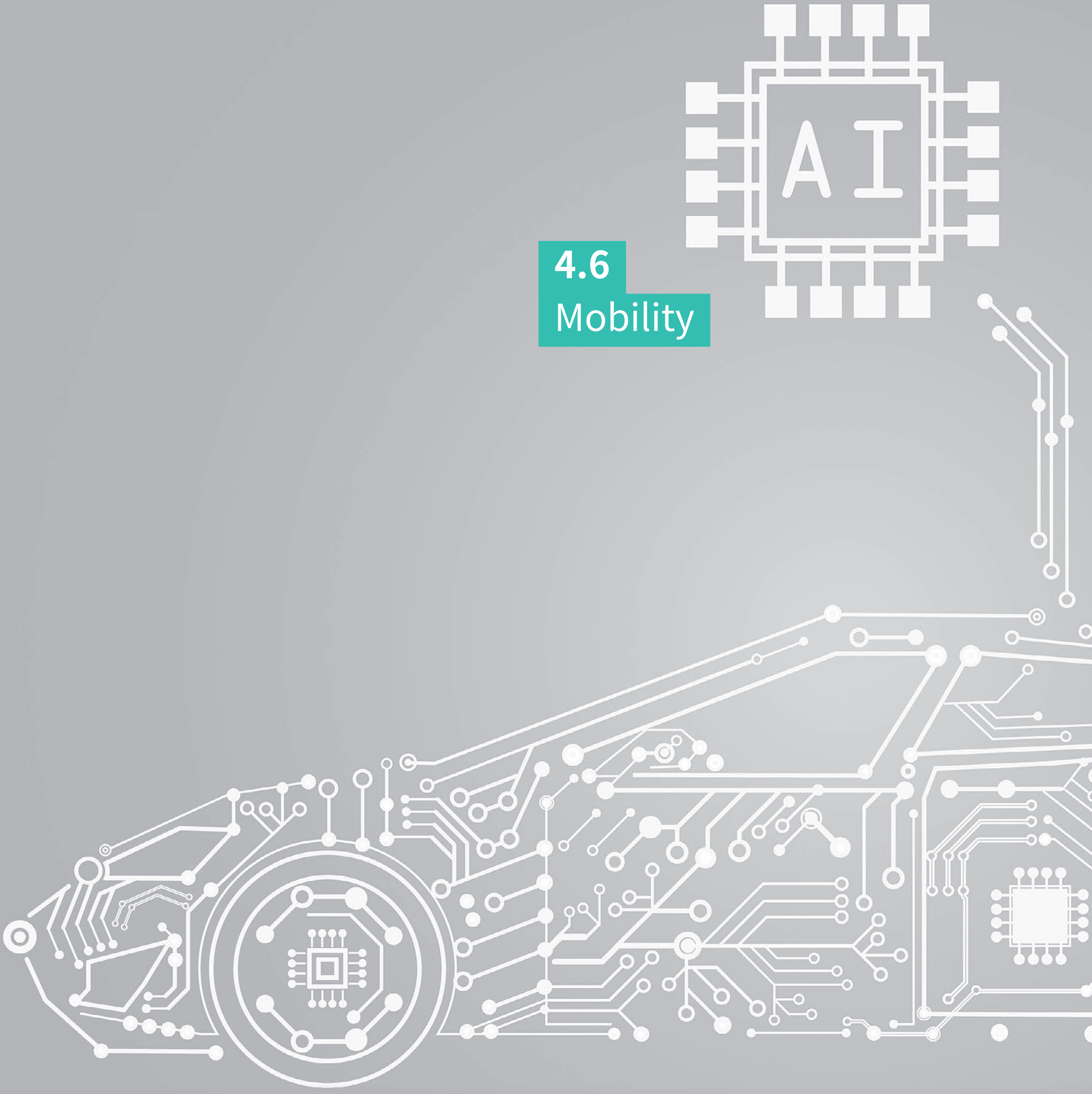


Urgency of implementation

Need	short-term			medium-term			long-term		
Standardization									
05-01		■							
05-02	■								
05-03					■				
05-04						■			
05-05							■		
05-06				■					
05-07				■					
05-08				■					
05-09					■				
05-10							■		
05-11					■				
05-12								■	
05-13				■					
05-14				■					
05-15								■	
05-16						■			
05-19								■	
05-20			■						
05-21			■						
05-22		■							
Research									
05-17									■
05-18						■			
Policy									
05-23			■						
05-24			■						

Figure 40: Prioritization of needs for the key topic Industrial Automation (Source: Working Group Industrial Automation)

4.6
Mobility



The mobility sector plays an outstanding role in terms of both its economic and overall social significance. Mobility is an essential factor in many important life decisions, enables participation in social life, and the transportation of people and goods is a basic requirement for a functioning economy. In addition, automotive engineering is still the industry with the highest sales and is an important employer in Germany.

On the one hand, the use of AI as a key technology offers important opportunities for the mobility sector, among other things by enabling complex automated driving functions and the optimization of traffic flows or complex mobility chains; on the other hand, it represents an enormous challenge, among other things because a safe and trustworthy use of AI requires far-reaching efforts in research, development, standardization and regulation. The transformation of the mobility sector through the use of AI is relatively advanced; among other things, many automated driving functions have already found their way into series-produced vehicles, and considerable sums are being invested in corresponding research and development (R&D).

In view of the high significance of mobility and AI in mobility, the focus in the following chapter is on comprehensively highlighting the current status, requirements and challenges, as well as the standardization needs in this sector. In contrast to the first edition of this Standardization Roadmap, which focused on the legal framework and the transport of goods (logistics), this second edition focuses on the following aspects:

1. Use of trustworthy artificial intelligence (TAI) in the mobility application domain and here, in particular, in the context of „cooperative, connected and automated mobility“ (CCAM). CCAM covers vehicles of different modalities (road, rail, water and air) with automated functions and their networking with intelligent infrastructures, such as in intermodal mobility.
2. Relevance of individual aspects of trustworthiness (cf. in more detail the chapter „Embedding and life cycles of AI systems“ below) in the context of system embedding on the one hand and the different life phases of the CCAM system on the other (cf. in more detail the chapter „Embedding and life cycles of AI systems“ below) as well as the related status of operationalization or operationalizability. Aspects of TAI considered include safety, IT security, robustness, performance, explainability, traceability, and human-machine interaction. In this context, safety in all mobility applications is of particular importance under the TAI aspects: It is „non-negotiable.” The requirements for the other TAI aspects depend on the application: Depending on the application context and valid regulations, certain minimum qualities must be realized here as well; however, the concrete characteristics of individual properties may well differ between different applications or the same applications of different manufacturers – also deliberately for product differentiation – as long as the high minimum qualities are met for each property. For safety, however, this must always be implemented „fully“, i.e. at the best level of implementation according to the respective state of the art, which demonstrably reduces the residual risk of damage occurring to below a minimum, socially accepted residual risk. Since the different TAI aspects are not independent of each other, there are contextual dependencies. In certain applications, for example, IT security can be an indispensable prerequisite for safety [311], [312] and must therefore meet equally strict requirements.
With this in mind, this chapter considers trustworthy AI from the following two perspectives:
   a) First, trustworthiness for AI systems in general, namely all TAI aspects are considered equally (the safety aspect is only highlighted insofar as it is necessary for explaining the overall context and classification in relation to the other TAI aspects). For this purpose, the three domains of highly automated driving, mobility services or chains, and infrastructure are considered or compared with each other in terms of specific functionalities as use cases.
   b) In addition, against the background of the special role of safety and the risks to life and limb of various road users inherent in the operation of CCAMs as described above, the aspect of safety is subjected to more in-depth consideration, particularly with regard to the verifiability of this property, as is necessary, for example, for type approval or certification of vehicles in the various application domains. Here, a differentiation is made between the modalities automotive, aviation, maritime and rail.

For social or ethical aspects for which specifications cannot be made purely from a technical point of view, the technical prerequisites necessary for controlling and enforcing these specifications are dealt with as examples, but the social implications are not addressed.

Trustworthy-AI-relevant aspects of and perspectives on AI systems

AI technology, including machine learning (ML) such as deep learning (DL) for deep neural networks (DNN), has become an indispensable key technology for many application areas, assisting humans in decision-making processes or even performing decision-making processes without human intervention. Trustworthy AI, i.e., AI that is trusted by people, organizations, and/or societies, is not only generally desirable, but is a prerequisite for the use of AI in safety-critical applications [313], [314], [315], [316], [317]. Whether an AI system possesses this "trustworthiness" property depends on the specific AI system, the specific application, and other framework conditions such as the legal and technical requirements for the development and use of this system [311], [312], [318], [319].

There are a variety of different perspectives on AI systems and TAI-relevant aspects, from which a large number of decisive criteria can be derived [312], [320]. These aspects and perspectives can have a (primarily) technical or social background. While the technical perspectives lead to purely technical criteria, the societal perspectives require technical foundations (in particular suitable metrics such as for the balance of datasets), yet the concrete requirements for this cannot be determined exclusively from the technical perspective, but require an ethical evaluation or socio-political classification. With regard to such criteria with (primarily) societal criteria – namely the acceptance of an AI system by individual users or society, fairness or bias, and data protection and privacy – this chapter focuses on the technical perspective, namely on the technical basis for verifying and enforcing these criteria (e.g.,: How can we support the testing of ethical criteria? Which metrics are particularly suitable for this?

Aspects of TAI that are particularly relevant in this context include:

→ Performance: Performance of the AI system in terms of relevant performance metrics.

→ IT security: passive and active robustness of the AI system against attacks and in particular against AI-specific attacks (adversarial attacks, „poisoning“ attacks and „privacy“ attacks) in relation to the three security objectives integrity, confidentiality and availability [83], [320].

→ Functional safety: A system is „safe“ if its operation does not pose any unacceptable risks to the environment (individuals, environment, organizations and goods).

→ Robustness and generalizability: Passive and active robustness to natural variations in inputs (situations) including those that could have been avoided if adequately accounted for during training. This includes robustness to stochastic influences such as noise and to interfering signals such as interference.

→ Explainability: Features of an AI system that enable humans to understand the AI system's decision-making process, either through inherently interpretable models or through post-hoc interpretation.

→ Interpretability: Properties of an AI system that make it possible for its performance to be monitored in the overall system. For this purpose, it is necessary to provide information for the plausibility of the results, which does not necessarily represent an „explanation“ in the sense of explainability.

→ Transparency, accountability and documentation (traceability): Traceability of the AI system throughout its life cycle, e.g., design decisions, constraints, data, models, training algorithms, training processes, evaluations, and operation, among others, through technical documentation and logging.

→ Risk management: Identification, analysis, and prioritization of risks and coordinated use of resources to minimize risk probabilities or risk impacts (acceptable marginal risk).

→ Human-machine interaction / "human oversight": Implementing human-in-the-loop/on-the-loop solutions – these can be seen as measures to increase safety or increase user engagement.

→ Acceptance by individual users and society.

→ Bias, impartiality: Measures to prevent unbalanced operation of AI systems, e.g., through training datasets that do not meet the IID criteria („independent and identically distributed“) and lead to discrimination, e.g., with respect to gender [317].

→ Data protection and privacy: Appropriate handling of sensitive (private and confidential) data.

→ Redundancy: What are the requirements for redundant data collection and evaluation in order to trust the overall system (also depending on the criticality of the functions), especially if black-box AI approaches (lack of interpretability or explainability) must necessarily be used because conventional algorithms cannot map the functions?

On closer examination, it becomes clear that the above-mentioned views cannot be sharply demarcated from one another and that there are numerous interdependencies. For example, there are overlaps between IT security and functional safety, since a successful security attack changes the functionality of the system and thus jeopardizes the functional safety of the system, and the non-fulfilment of a functional safety property can open up attack surfaces for security attacks. Lack of robustness due to built-in semantic plausibility, for example, makes a number of attack patterns more likely or even possible in the first place. These include "adversarial attacks" by pixel-level manipulation, which can be intercepted in TAI (cf. [321]). The two properties are thus mutual prerequisites for the overall consideration of the system. Other interdependencies between TAI aspects obviously exist as well; central to this Roadmap is the finding that a large number of aspects are highly relevant to TAI systems and must be considered accordingly. The relevance and the necessary prioritization of the respective aspects must be evaluated separately for each application or application class.

Integration and life cycles of AI systems

Systems referred to as AI systems (understood here in terms of, among others, the draft ISO/IEC 22989:2022 [16]) often consist of multiple interacting software and hardware modules and are embedded in an overall system of AI and non-AI components and in relation to a context [318]. For example, the software system consists of a variable number of classical IT modules, symbolic AI modules (e.g., logical reasoning or decision trees), and connectionist AI modules (e.g., neural networks) that communicate with each other via suitable interfaces. The software runs on computing units, each of which can be connected locally (edge) or via a network (cloud). The software interacts with the environment via hardware modules. For example, an automated vehicle has a large number of sensors and actuators connected via mechanical, electrical and IT systems in the vehicle "body". Sensors can be divided into proprioceptive (internal sensors such as wheel rotation sensors), exteroceptive (external sensors such as camera sensors), and virtual sensors (such as inputs from communication channels or from the fusion of different sensors). Actuators range from the powertrain to the braking system and steering system to the lighting system and user-relevant information systems (display, loudspeaker). The environment of an AI system here can be various passive and active traffic participants, occupants of the embedding vehicle, or smart city infrastructures, among others. There is usually an organization behind the development of an AI system, and one or more organizations also have responsibilities during the operation of the AI system, due to the provision or processing of data streams. Therefore, these organizations must also be included in the overall consideration. Overall, the application-specific embedding of an AI system results in application-specific requirements and risks that should be considered when developing, testing, and operating such a system.

In classical IT and symbolic AI systems, the structure and parameters can, at least in principle, be defined or set directly by the developer, and their functionality can be reproduced in operational use. However, for IT and symbolic AI systems above a certain critical size, the large number of parameters may make it difficult or impossible to directly design and tune the parameters and interpret how they work. In connectionist AI systems such as neural networks and support vector machines, this problem is much more pronounced due to the fact that their processing is not (a priori) intuitive for humans, and thus applies to a large part of the systems in use. Such systems must be developed in a data-driven, iterative training cycle using machine learning techniques, in which the developers define the framework in each case, but no longer directly define the parameters of the operational system. This results in a complex life cycle, which can be divided into the following phases, as has become apparent in practice:

→ Planning phase: Here, depending on the desired characteristics of the AI system to be developed, suitable AI models, learning methods, required data, metrics, and quality assurance measures, including any dependencies, are identified, among other things, and a development plan is defined.

→ Data collection and QA Phase: Data required for training is obtained in sufficient quality and quantity. Here, data can initially be self-obtained (data acquisition in the physical world), obtained from external sources or generated synthetically. In addition to combining these data sources, data can be enriched in a variety of ways, for example, to increase the number of data or to incorporate desired properties into the dataset. Depending on the specific requirements of the dataset, various quality assurance measures follow.

→ Training phase: The developer iteratively starts one or more training processes with predefined models, data and hyperparameters. Depending on defined termination criteria, which are checked using appropriate metrics (e.g., performance criteria), the training processes are stopped and restarted with adjusted parameters until at least one trained system meets the predefined requirements (regarding the defined metrics and quality assurance measures).

→ Evaluation phase: Evaluation of the system goes beyond the automated calculation of metrics in the training process and is performed before, during, and after the system goes live. For the evaluation e.g. complex simulations or pentests[92] can be used

→ Deployment and scaling phase: Here, the AI systems are adapted for practical use and commissioning, which may include further optimizations, e.g. with regard to improved scaling or efficiency.

→ Operational phase, including maintenance: In principle, it would be conceivable to carry out further training phases in the operational phase as well (so-called self-learning or online learning systems). However, the resulting changes in system behaviour are completely beyond the scope of a safety analysis and the associated safety verifications compared to the current state of the art, so that certification or type approval of such systems is currently not possible. This type of AI system is therefore not considered in this chapter.

→ Retirement: If the AI model and/or training data are to be protected against privacy attacks on the model and/or data even after regular operation (e.g., for privacy or IP reasons), an orderly decommissioning that permanently prevents public access to the model and data is required. Otherwise, this life cycle phase has no AI-specific relevance.

Due to changing requirements, due to weak points of the system becoming known during operation, or due to the goal of continually improving a system, the above-mentioned phases are run through cyclically (continually in the sense of a continual development process). There is a continuous transition ranging from infrequent, carefully planned and executed updates with, if necessary, significant changes to the previous version, through to very short update cycles, and to self-learning or online-learning systems. While discrete updates are now indispensable for many systems and are carried out regularly, self-learning systems (i.e., systems that adapt in the field based on incoming observations) have not yet been used in safety-critical applications such as mobility, despite a great deal of media attention (see also above regarding certifiability).

4.6.1 Status quo

4.6.1.1 Fundamental, qualitatively novel properties of AI technology

On the one hand, the use of AI technology opens up new opportunities and enables applications that cannot be realized with classic technologies, or are realizable only to a very limited extent. On the other hand, the complexity of AI systems and their life cycles leads to qualitatively new problems and risks [320], [83]. As described above, the development of AI systems usually requires a data-driven approach, and the developer has no direct control over the learned parameters of the AI system and the input/output correlations implied by them. As a result, operational AI systems have black-box properties, and their modes of operation (and thus also possible errors) are not directly apparent to the developers and users. The properties of the functions implicitly encoded using machine learning and data depend significantly on the underlying training dataset. However, sufficient quality assurance of training data is a non-trivial task, especially when the data come from external sources. If pre-trained models are used, as is often practiced, there may be hard-to-detect vulnerabilities in the AI system that often survive further post-training sessions unscathed. Many AI systems also have a huge input and parameter space. The camera input of a 4K camera with a high number of colour channels can be mentioned here as an example. As a result of this complexity, formal verification methods are not available for many practically deployed AI systems, and alternative empirical validation methods can only cover a fraction of the parameter space for practical reasons. Thus, an AI system does not necessarily fulfil the programmer's intent, and there is no guarantee of what has been learned by the system, nor any certainty regarding the trustworthiness aspects listed at the beginning (cf. Chapter 4.6) – such as what performance the system will achieve in practice. Conversely, there is often no or only a limited explanation of how an AI system works for humans. Regarding the various trustworthiness aspects of AI systems, the technical understanding is currently incomplete, including functionality, integrity, reliability, safety, and generalizability, and further extensive R&D efforts are needed.

[92] Penetration tests, i.e. controlled cyber attacks with the aim of identifying vulnerabilities.

4.6.1.2 Requirements, testing and safeguarding of AI systems

The increasing use of AI technologies on the one hand, especially in safety-critical systems, and the high complexity of AI systems on the other hand, which leads to qualitatively new risks, result in increasing needs for regulation, standardization, safeguarding and objective verifiability of AI systems with regard to their trustworthiness. Initial approaches in this direction, especially at the abstract level, already exist, such as the planned horizontal (i.e., cross-sectoral) regulation of AI systems in the European Union (EU), the AI Act [4]. Despite major international efforts in R&D, there is currently a lack of sufficiently practical and technically sound requirements, testing and mitigation strategies and the corresponding tools [312]. Since formal verification methods are often practically not applicable due to the system complexity, empirical validation and testing methods have to be used. However, sufficient validity here requires very good coverage of the input space of the system. In order to achieve sufficient test coverage, especially including relevant corner cases, it may be necessary to restrict relevant boundary conditions, in addition to technical developments. This can mean, for example, limiting automated driving functions to certain traffic situations and weather conditions.

In any case, at this point in time, the requirements regarding the relevant aspects have not been specified comprehensively – across the entire life cycle of an AI system. Corresponding metrics, which – analogous to the concept of “key performance indicators” (KPI) – can function as “key trustworthiness indicators” (KTI), have not yet been sufficiently established. Due to the complexity of the topic, it is advisable to focus on specific application classes and applications. This vertical approach, which is complementary to the horizontal approach of the European Artificial Intelligence Act (AI Act), has the medium-term goal of operationalizing the AI Act initially for individual applications and the long-term goal of generalizing the sector-specific findings and iteratively improving the horizontal model.

In addition to improving the testability of AI systems with respect to their TAI properties, another key goal is to develop AI systems from the ground up so that they possess essential TAI properties (“trustworthy by design”).

4.6.1.3 State of the art, current use cases

Automotive

In the automotive sector, the use of AI is currently mostly limited to non- or limited safety-critical functions or prototypes without series approval. In addition to driver assistance systems, which support human drivers in certain driving situations, systems with a higher degree of automation are only used in very defined areas [311]. Such systems with defined operating ranges include automated valet parking (AVP) systems and automated lane keeping systems (ALKS). An ALKS essentially takes over the longitudinal and lateral guidance of a vehicle. This system may only be used up to a speed of 60km/h in special areas (operational design domain, ODD), in which a structural separation of the driving directions prevails and which are closed to particularly vulnerable road users such as pedestrians and cyclists under normal circumstances. In Germany, this essentially applies to parts of the federal highways (especially motor roads) and motorways. In this area, the system takes over both the adjustment of speed and steering to follow the road layout within the “open spaces” in the lanes used by the vehicle. The sensor technology of the system takes over the recognition of the “open spaces” in the environment. Monitoring of the function takes place to a limited extent through the system itself (e.g., detection of interference by or absence of a vehicle driver(s) as well as technical difficulties in maintaining the lane, speed or determining the “open spaces”). As long as the system is activated, it is in charge, but the system must permanently ensure that vehicle drivers can take over the driving function within a certain time frame (e.g. ten seconds). If a requested handover does not occur, the system must perform what is called a minimum risk manoeuvre (to reach a state where the risk is minimal). The world’s first approved system of this kind is the “Drive Pilot” system from Mercedes, which complies with the SAE (Society of Automotive Engineers) Level 3 automation grade. An extension of the regulations, which among other things will allow driving up to a speed of 130km/h and lane changes, has already been prepared and will come into force at the beginning of 2023.

In addition, national regulations apply, such as the Gesetz zur Änderung des Straßenverkehrsgesetzes und des Pflichtversicherungsgesetzes – Gesetz zum Autonomen Fahren (Act Amending the Road Traffic Act and the Compulsory Insurance Act – Autonomous Driving Act) of July 12 [322] and the associated Autonome-Fahrzeuge-Genehmigungs-und-Betriebs-Verordnung (Autonomous Vehicles Approval and Operation Ordinance) [323].

The most important standards and specifications on which type approval is based are the ISO standards on functional safety (ISO 2662 series [455]), cybersecurity (ISO/SAE 21434:2021 [324]) and on functional safety of the intended function (SOTIF, ISO 21448:2022 [90]). More recent standards, some of which are still under development, extend these to include concepts of scenario-based testing, in-service testing, and consideration of highly automated driving functions, such as those enabled by the use of AI processes (ISO/TR 4804:2020 [325]) and its successors ISO/TS 5083 [326], ISO 22737:2021 [327], ISO PAS 8800 [110].

Aviation

In the field of aviation, a distinction can first be made between different application domains. In this context, the field of urban air mobility (UAM) is clearly distinguished from the field of conventional aviation. Through the UAM and especially the development of drones and air cabs, many new actors are involved which are characterized by their short development cycles and strong affinity to technology. The potential areas of application for AI are diverse in both application domains, although there are still no clear standards and specifications for AI-based functions due to the generally high safety and certification requirements and the redundancy of systems generally required in aviation. Although there is already an AI roadmap from EASA [328], see Figure 41, as well as initial concrete work on specific "concepts of operations" and considerations of the trustworthiness, explainability, and reliability of AI [329], [330], the same, if not greater, hurdles arise in translating the set goals into certification procedures and standards compared to the other domains addressed in this chapter. In principle, the automation steps are oriented to the transfer of responsibility to AI-based functions. The introduction of pilot assistance systems is addressed first. Subsequently, single-pilot operations are to be implemented, in which the AI merely supports the pilot. Thereafter, responsibility is to be gradually transferred to the AI, so that first partially automated functions and finally fully automated functions/aerial vehicles are implemented.

The first AI applications in the field of pilot assistance systems (for example, Deadalean and Iris Automation) and small drones are already established due to the manageable risk (low weight, pilot for monitoring); here, drones follow individual persons, map structures automatically or are capable of independently recording their surroundings in three dimensions (for example [331]). However, even at this scale, there are still no meaningful ways to enable certification of procedures, which underscores the need for specific recommendations for action to create such standards and specifications.

The fact that many of the aerial vehicles (AV) can cause significant damage to the AV itself, other AV (air risk) or the environment (ground risk) due to a incorrect behaviour of relevant AI functions is problematic in the area of aviation. The criticality increases additionally especially when people are passively or actively involved. Thus, the risk associated with aviation automation will need to be assessed, not least because of the kinetic and potential energy of an unmanned aerial vehicle (UAV), as well as the involvement of people, for

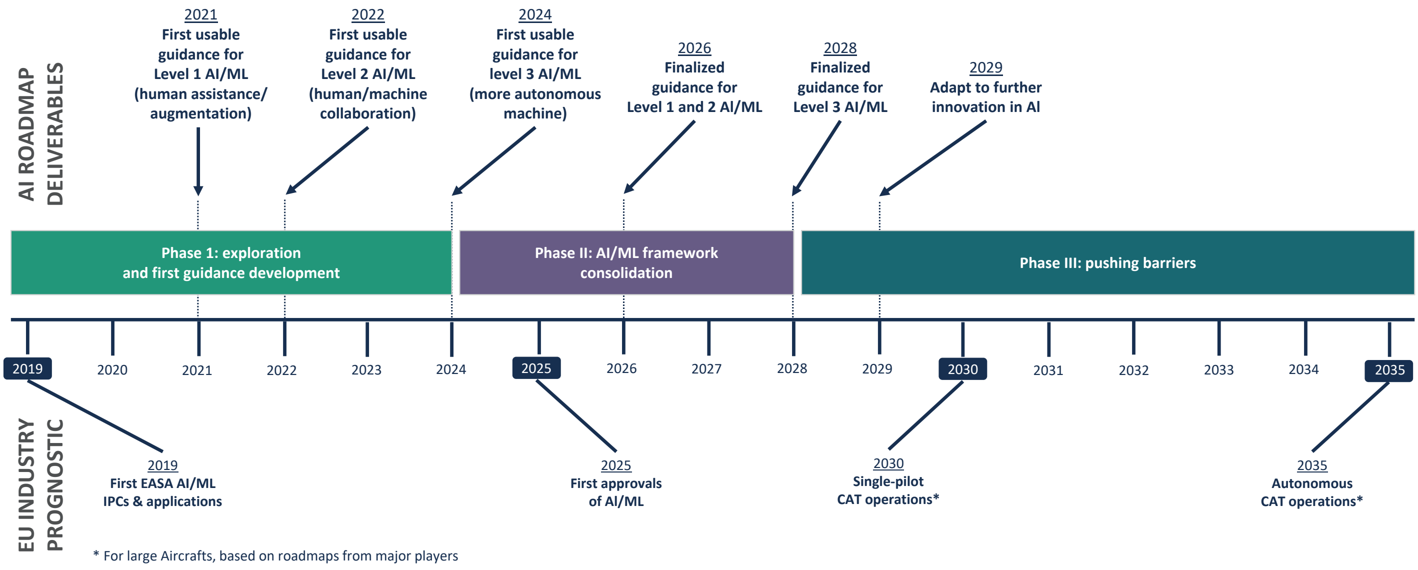

AI ROADMAP DELIVERABLES
2021 First usable guidance for Level 1 AI/ML (human assistance/ augmentation)
2022 First usable guidance for Level 2 AI/ML (human/machine collaboration)
2024 First usable guidance for level 3 AI/ML (more autonomous machine)
2026 Finalized guidance for Level 1 and 2 AI/ML
2028 Finalized guidance for Level 3 AI/ML
2029 Adapt to further innovation in AI
Phase 1: exploration and first guidance development
Phase II: AI/ML framework consolidation
Phase III: pushing barriers
2019
2020
2021
2022
2023
2024
2025
2026
2027
2028
2029
2030
2031
2032
2033
2034
2035
EU INDUSTRY PROGNOSTIC
2019 First EASA AI/ML IPCs & applications
2025 First approvals of AI/ML
2030 Single-pilot CAT operations*
2035 Autonomous CAT operations*
* For large Aircrafts, based on roadmaps from major players

Figure 41: Excerpt from the EASA Artificial Intelligence Roadmap (Source: along the lines of [328])

the various classes of AV and their respective areas of operation. Another crucial factor is, of course, the trustworthiness and reliability of the AI, as well as the criticality of its area of application for the functionality of the AV.

Due to the high risks associated with transporting goods or people across the third dimension, work on integrating AI is generally much less advanced than in the domain of highly automated driving, for example. Nevertheless, future use cases can be outlined, especially with regard to safety-critical functions.

Maritime

In maritime applications, systems with AI components or standalone AI systems are increasingly being used, for example, to support nautical staff in decision-making. These are often optional equipment variants or special functions that are offered as an additional feature alongside the equipment that must be equipped in accordance with the SOLAS (International Convention for the Safety of Life at Sea) or the MED (Marine Equipment Directive) and are not subject to any specific approval. These include, for example, 360-degree perception systems that capture the environment and display it on a separate screen with annotated AR elements. In addition to classic collision prevention systems (e.g. radar, AIS), data-based collision warning systems are also used. Such assistance systems only serve to provide information and do not trigger any independent actions. If necessary, automated suggestions will be made. Thus, the human is still the controlling authority and the ship's command bears the responsibility for decisions made. The ship's bridge must therefore be permanently manned by qualified personnel. Systems that go beyond this are currently prototypes or part of research projects and are used for testing under controlled boundary conditions, with humans also acting as controllers to intervene in dangerous situations.

There are four different degrees of automation. For this purpose, the International Maritime Organization has defined the Maritime Autonomous Surface Ship with the associated degrees of automation:

→ **Degree One**
Ship with automated processes and decision support: Seafarers are on board to operate and control shipboard systems and functions. Some operations may be automated and at times be unsupervised but with seafarers on board ready to take control.

→ **Degree Two**
Remotely controlled ship with seafarers on board: The ship is controlled and operated from another location. Seafarers are available on board to take control and to operate the shipboard systems and functions.

→ **Degree Three**
Remotely controlled ship without seafarers on board: The ship is controlled and operated from another location. There are no seafarers on board.

→ **Degree Four**
Fully autonomous ship: The operating system of the ship is able to make decisions and determine actions by itself.

Degree One can already be assigned to many new ships. Control systems such as autopilots take control of the drive under human supervision and follow a set route. Such systems are rule-based and do not require AI.

In the field of shipping, there are currently no standards or specifications focusing on AI components.

Railway

Railroads differ from automated urban rail-based passenger transport (AUGT) according to DIN EN 62267:2010 [332], such as the Nuremberg Railway [332], in terms of the absence of safe barriers. Rail traffic takes place in the open air and thus implies the presence of non-system obstacles. DIN EN 62267:2010 [332] lists grades of automation from GoA0 to GoA4 (see Table 9). There is not yet a standard that classifies grades of automation for the railway sector. However, a classification similar to DIN EN 62267:2010 [332] is generally used. System-internal obstacles such as other rail vehicles are indicated to the train drivers in good time by signals starting from GoA1. GoA0 is also referred to as on-sight train operation. Additional visual detection of other rail vehicles is therefore not necessary with GoA1. The automation grade GoA2 again means that the train driver is responsible for detecting non-system obstacles, opening and closing doors and emergencies. GoA3 means that the train crew is only responsible for emergencies and can move freely in the train. GoA4 describes trains without a train crew. In GoA4, a monitoring and control centre staffed by humans can be used. Shunting yards are an exception in GoA4 – other rail vehicles must also be detected visually there. GoA2 is currently the state of the art. Grades from GoA3 and above are still experimental, as for example in the case of "AutoHaul" [334].

Table 9: Simplified overview of automation grades for railroads

	GoA0	GoA1	GoA2	GoA3	GoA4
	On-sight train operation	Driving by signals	Semi-automated train operation	Driverless train operation	Unattended train operation
Switching and driving authorizations	X	S	S	S	S
Distance between trains	X	S	S	S	S
Braking	X	X and S	S	S	S
Acceleration	X	X	S	S	S
Obstacle detection	X	X	X	S	S
Supervising passenger transfer	X	X	X	X or S	X
Monitoring status	X	X	X	X	S
Emergency situations	X	X	X	X	S and/or operations control centre

X – operations staff, S – technical system

As can be seen in Table 9, obstacle detection is the key challenge for GoA3. Experimental systems for obstacle detection in the railway sector have been tested since the 1990s [335], and the implementation often uses conventional image processing technologies. The use of AI systems for obstacle detection is often considered a more powerful solution. AI models and methods are increasingly being applied to diagnose and predict remaining service life, fault events, or other status characteristics of rail infrastructure elements. These innovations form the basis for the development of digital and data-based maintenance strategies such as status-based and predictive maintenance.

4.6.2 Requirements and challenges

As described in the previous chapter, safety occupies a special position among trustworthiness aspects. The following Chapter 4.6.2.1 therefore first looks at the trustworthiness property as a whole, while the next Chapter 4.6.2.2 focuses on the safety aspect and presents the other challenges that apply specifically to this trustworthiness aspect.

4.6.2.1 Trustworthy AI-based mobility

The trustworthy use of AI technology in mobility is complex and determined by various dimensions of complexity (cf. Chapter 4.6.1). Due to the limited resources available, the multidimensional space thus spanned can only be discussed here in the form of selected application examples, nevertheless with the aim of developing application-specific requirements and challenges as justification for the needs formulated in Chapter 4.6.3. With the objective of selecting fewer use cases in such a way that they together cover the complex space of trustworthy AI in mobility as well as possible, and correspond to the expertise available in Working Group Mobility for a sufficiently in-depth consideration, the following use cases in the three different domains were selected for further consideration:

1. **Evasive manoeuvres** as a complex driving manoeuvre in automated driving
2. **Ride Sharing** as part of mobility services
3. Traffic optimization via an improvement of the **traffic signal control in the traffic infrastructure**

For all three application examples, the relevance and then the status of the operationalization of the planned regulation of AI in the respective application example were first systematically recorded in order to then be able to compare and generalize them. A modified version of the Certification Readiness Matrix [312], referred to as the Trustworthiness Readiness Matrix (TRM), was used for the systematic survey. The TRM maps the following two dimensions:

→ Embedding phases (organization, application-specific requirements and risks, embodiment and situatedness of the AI mode) and life cycle phases (planning phase, data collection and quality assurance phase, training phase, evaluation phase, deployment and scaling phase, operational and maintenance phase).

→ Trustworthiness aspects (safety, security, performance, robustness, interpretability/explainability, traceability/documentation/logs, fairness/impartiality, data privacy)

More detailed results in the form of the TRMs can be found in Annex 13.4 "Trustworthiness Readiness: Selected Results". The most important results are summarized in compact form below.

By comparatively contrasting TRMs for different application domains/areas (for example, automotive vs. medical) and AI technologies (for example, decision trees vs. DNNs), additional dimensions could be considered and tracked over time.

Use case evasive manoeuvres as complex driving manoeuvres in automated driving

Automatic driving functions are probably the best-known application of AI in mobility and – in view of the unpredictable environmental conditions (besides traffic itself, weather conditions, road conditions, sensor pollution, unrecognizable road signs, etc.) – one of the more complex applications. The high complexity of this application, combined with high risks of harm during operation and the high and immediate application relevance based on the already extensively advanced automation of many driving functions suggests a consideration of this application. Since automated driving itself encompasses a very wide range of scenarios and functions, the use case under consideration was further narrowed down to the functionality of the evasive manoeuvre as a complex and representative driving manoeuvre within automated driving (SAE Level 2 and 3, Advanced Driver Assistance Systems).

Namely, the present use case consists of the task of the AI system to plan an evasive manoeuvre based on the output of the sensors and to execute it by giving appropriate commands to the vehicle to control the direction of movement and speed. Planning involves deciding both whether and how to perform the planned actions, which includes complexities such as safety distance, delay in the flow of information (both from the incoming signals from the sensors and from the communication with the vehicle's control system), speed adaptation and braking distance calculation. Furthermore, with regard to decision-making, not only the aspect of safety (i.e., avoiding a collision) must be considered, but also efficiency (i.e., the duration of the manoeuvre) and driving comfort (i.e., avoiding sudden changes in speed and motion). Furthermore, it was decided to focus on the perception level (object detection) including sensor function.

1. A systematic examination of the use case with the help of the TRM with regard to the relevance of the trustworthiness aspects in the respective life cycle and embedding phases shows the consistently high relevance of all aspects: The safety aspect, which is particularly important in the operational phase, plays a prominent role here: People and the environment must not be harmed. Against the same background and in view of the fact that an attack on the system by external parties must be ruled out, the security aspect was rated as highly relevant. The aspect of performance has a similarly high relevance – also as a prerequisite or support for safety – which also has a very high relevance in the operational phase, as well as robustness, and this tends to be in the later life cycle phases. Here, the understanding was taken that performance means an average value for all cases/situations occurring, while robustness is understood as the reliability of the system even under predictable „extreme" („edge cases") and unpredictable („corner cases") conditions. It is particularly important to take such corner cases into account in evasive situations, as these are by their very nature unusual situations in road traffic. Corner cases frequently emerge in the present use case not least because of the inaccuracies/discrepancy between the simulated environment in which the AI system is trained and the real environment in which it is deployed. In this respect, a standard must be developed on which threshold values must be reached for which concrete factors/functions in order to be able to speak of robustness or to guarantee it. Data privacy was assessed as highly relevant at points in the operational phase, particularly in relation to the issue of the extent of storage of sensor data for traceability, which regularly includes image data from other vehicles and people. Furthermore, a high relevance was seen for interpretability/explainability – namely in the evaluation, deployment and operational phases – as well as for

traceability – namely in addition to the operational phase under the rubric of embedding for the application-specific requirements and risks. The background to this is in particular the clarification of the question of guilt in the event of an accident (both in the event of an accident despite an evasive manoeuvre and in the event of an accident due to an evasive manoeuvre). Fairness/impartiality, which plays a role especially in public discussions of evasive manoeuvres in relation to (extreme) conflict situations, is considered particularly relevant for the data collection and evaluation phase.

2. Accordingly, while the overall relevance of the entire trustworthiness complexity space for this application is high to very high, the state of operationalization of regulation is more „mixed." There are now many standardization initiatives (including [336]; ENISA Ad-Hoc Working Group on AI Cybersecurity; NIST Trustworthy and Responsible AI; Project AI Safeguarding; Grand Défi – Sécuriser, certifier et fiabiliser les systèmes fondés sur l'intelligence artificielle; UNECE GRVA technical workshop on Artificial Intelligence). Aspects that can be served by existing methods, e.g. traceability, are likely to be operationalizable with comparatively little effort. However, especially for the technical aspects of safety, IT security and robustness, an operationalization of the planned EU AI Act is not yet possible [312]. After all, the introduction of such highly and fully automated driving functions is only considered justified if an improvement in the safety and environmental compatibility of road traffic is also demonstrated. In order to bring such driving functions quickly into circulation, in addition to the regulation of (AI) technologies, a reform of type approval should be sought with regard to their dynamization.

Use Case Ridesharing as a mobility service (mobility chain)

The flexible and time-limited allocation of vehicles to customers by commercial providers in the context of "ridesharing" falls under the class of "mobility services" such as "mobility chains". On the one hand, AI is used here in the context of automated driving functions, where – unlike in regular private transport – a wide variety of users interact with a wide variety of vehicle types, including different levels of automation, under shorter familiarization periods. On the other hand, AI is used to optimize fleet management including vehicle provisioning and allocation, as well as predictive maintenance.

These services are based on applications and vehicles with automation functions, the correct operation of which requires an appropriate level of trust on the part of the user. Accordingly, only the additional challenges compared to automated driving will be considered here with regard to the vehicle (i.e., the requirements assumed for automated/autonomous driving – such as for safety and security – are assumed as a basis). The use case considered here is the most complex case, in which a mobility provider has vehicles on offer from different manufacturers with different automation functions, operating concepts, etc.

1. A systematic consideration of the complexity space with the help of a TRM with regard to relevance in relation to the use case shows that the following additional challenges in particular arise in comparison to automated driving:
   a) AI functions and implications of the vehicles must be sufficiently explained to the users (individuals and organizations) within a short time, so that an increased relevance must be placed on the aspect of explainability.
   b) The functions are to be designed for very heterogeneous users/environmental profiles, whereby functions should also be deactivated or deactivatable if necessary, depending on the user's prior experience or user needs. This applies in particular insofar as the aspects of safety and security require it.
   c) In the case of "ride hailing", the allocation of vehicles to users should be based not only on economic aspects, but also on fairness principles. This is of particular relevance insofar as private mobility service providers will basically operate according to economic efficiency or profit aspects. At this point, any incentivization on the part of the public sector will be an important question to be answered politically. These aspects have an overriding relevance for the performance of such mobility services/chains.
   d) In this context, it will also play a role to what extent / at which locations the public sector provides interfaces between the different modes of transport – for example, drop-off zones with access to the public transport network (rail or waterways) to be accessed by cars. The same applies to the question of the extent to which multi-layer traffic optimization should be promoted and (private) mobility providers should be integrated.
   e) Fleet maintenance, including predictive maintenance methods, should be based on uniform standards, even for different manufacturers and models.
   f) Since the aforementioned functionalities involve processing data from different vehicles with frequently

changing users, the aspect of data protection – in general and in the operational phase in particular – is of outstanding importance. The necessity of the extent of data collection is conditioned by the degree of automation of the vehicles, as well as the fleet control on the part of the mobility operators. Its acceptability depends, however, on the profile of the mobility user as well as on the purpose of use, so it can be assumed that the aspect of data protection in ride sharing cannot be regulated purely on a vehicle basis (see [337]).

g) It has been identified as essential that the responsibilities between vehicle manufacturers and mobility service providers with regard to the requirements specified by the legislator or standard setter in the public transport area – in particular, but not only, the requirements for safety and security – must be clarified in a binding manner and then also communicated transparently to the customer. This applies, for example, with regard to software updates, which can affect the vehicle applications on the one hand and the service application on the other, and in this respect in particular the interfaces between the applications (such as access by the service provider application to the vehicle‘s navigation system / original equipment manufacturer (OEM)). This makes the aspect of documentation more relevant.

2. With regard to operationalization, it should be noted that, compared to the status/executions in autonomous driving (see above section „Use Case ride sharing as a mobility service (mobility chain)“), operationalization of the aspects relevant for mobility services/chains has only been advanced in a few areas so far. While the data protection aspect can be based on the approaches to fleet management, no reliable approaches can be identified with regard to either the safety/security aspects mentioned or performance and explainability/fairness; especially since the AI-based developments have so far been driven forward in particular by OEMs, and the relevant know-how and embedding within the organization can therefore be observed in the latter, while they are currently hardly available among the mobility providers.

Use case traffic optimization via an improvement of the traffic signal control in the traffic infrastructure

The optimization of traffic flow with the involvement of various road users is of great importance in mobility. The control of traffic signals in itself is obviously highly relevant to safety. For this reason, the use of AI technology should currently only be used in parallel with or embedded in classic, formally verified processes that guarantee compliance with all safety-relevant aspects. In the future, however, AI-based control functions cannot be ruled out, e.g., in the context of smart cities, such as the inclusion of environment recognition, calculation of optimal phase sequences/transitions/durations and, in particular, transition times between red-green phases and, in particular, interaction with or recourse to V2X data (car-to-infrastructure data). Their influence on traffic safety should therefore be included in the considerations.

1. A systematic consideration of the complexity space using the TRM in terms of **relevance** with respect to the use case shows that the following challenges in particular exist when using AI in traffic signal control:
   a) Since personal data can (also) be used in the data collection phase (whereas the use of personal or anonymized data is sufficient in the deployment or scaling phase and the operational phase, or critical data can be processed on-chip without access to raw data or personal data), a high relevance for data privacy was assumed here. Against this background, the aspect of security – especially in the phase of data collection – was also considered highly relevant.
   b) The aspects of fairness and traceability of decision-making processes with regard to different road users (groups), modes of travel (passenger car, motorized two-wheeler, bicycle or pedestrian, etc.) and routes (e.g. main traffic flow to secondary flows) is rated as highly relevant, especially since this also has a direct influence on the performance of the infrastructure system. Fairness with regard to multimodal aspects (for example, weighting of different types of transport users in the definition of the target function/optimization variable) also plays a special role. The aspect of traceability is also considered to be extremely relevant in the deployment phase in that the „rolling out“ of a successfully tested system to various municipalities requires good documentation, especially since it must be possible for traffic engineers without AI expertise to operate the system.
   c) In addition, performance and robustness are considered to be particularly relevant aspects, especially in the case of high traffic volumes, bad weather, etc. This reflects the purpose of traffic optimization, according to which the traffic flow should be better controlled than today in general in normal/average cases as well as in extreme cases, and at peak times or in extreme situations.

d) Interpretability was seen to be highly relevant in the early stages of data collection, training, and evaluation. Since key decisions regarding design are made in the development phases, they must be justifiable. In addition, however, the AI‘s decisions must also be comprehensible to the users in the deployment phase, for which the foundation is laid in the system design, i.e. in the early development phases.

e) The relevance of safety and security is not particularly high – at least in comparison to the use cases considered above – because redundant classic systems are used here and data cannot simply be manipulated directly by the user. Nevertheless, the consideration of safety and security is also indispensable here.

With regard to the status of **operationalization**, it should first be noted that a legal framework for traffic control systems and the requirements and test criteria derived from it already exist. On the one hand, however, it should be noted that even in those areas where the legal framework has already been set quite comprehensively – especially for the area of data protection – operationalization in the sense of an operationalized requirements catalogue does not yet exist with regard to individual phases (for example, in the aforementioned area of data protection for the data collection and deployment phase). On the other hand, the need for operationalization is high with respect to future multidimensional or multimodal systems (such as, in particular, coordinated traffic signal control systems taking into account hierarchies as in the case of, for example, emergency vehicles, different areas at different times of the day or, for example, rush hour up to vacation times or special events) as well as multi-agent functionalities. This is all the more true since such systems/functionalities promote the performance and robustness that are classified as particularly relevant above. In this context, it should also be noted that currently for ad hoc situation detection – and thus especially in the deployment and operationalization phase – the state of sensor technology is insufficient.

Resulting challenges
Looking at the aspects with high relevance on the one hand and/or low operationalization level on the other, high to very high needs emerge almost across the entire TRM and thus for the entire complexity space. The needs range from the recording of the current status to the development of missing fundamentals, the formulation of contemporary requirements, the availability of recommendations for action, and the provision of a suitable infrastructure of scenarios, data, and simulations.

4.6.2.2 Safe highly automated mobility

Future use cases

Modality automobile
In the automotive sector, the ALKS represents the current state of the art (see Chapter 4.6.1); the highway chauffeur, automated hub-to-hub transport and automated driving in urban areas – in increasing complexity – with various intermediate stages are joining the ranks.

The **highway chauffeur** is essentially a further development of the ALKS. First, this is characterized by an expansion of ODD, so the highway chauffeur performs advanced driving manoeuvres such as overtaking other vehicles or changing highways at interchanges, and generally drives at higher speeds. However, the operation of the highway chauffeur is still only allowed in structurally separated, standardized areas that are not permitted for particularly vulnerable road users. Along with this expansion comes the need for more complex sensory awareness of the environment. This includes, in particular, the detection of objects and the prediction of the future behaviour of these objects. However, a rough classification of the objects (e.g. into static, dynamic, motorcycle, car, truck) and an estimate of the motion vector seems sufficient here. In this case, the driving function of the system is still redundant, so the monitoring of operational design domain (ODD) compliance in particular is taken over by the vehicle and control is actively transferred to the human when leaving the ODD, who must be ready at almost any time. Humans serve as a fallback level for the performance limits of the system, but they can actively take control themselves at any time. The potential use of AI in this use case includes sensor data processing in the system, sensor data fusion, object detection and motion prediction for environment awareness, tactical planning and trajectory planning including an assessment and monitoring of trajectories, and attention monitoring of the driver(s). Furthermore, AI could also be used in the area of development and testing of the systems to intelligently explore test cases, as well as identify corner cases. Like the ALKS, the highway chauffeur is an SAE Level 3 automation level system.

Unlike ALKS and highway chauffeur, automated **hub-to-hub transport** is not a convenience function for vehicle drivers. On the contrary, such transport is mostly fully automated and takes place without the presence of people in the vehicle. Compared to the highway chauffeur, the ODD in this case is extended to include construction sites and depots, and their entrances and exits with instructed personnel and is thus still structurally separated. Although the system within the ODD performs the same functions as the highway chauffeur, the presence of a human as a fallback level is eliminated. For this reason, the functions must be performed with higher quality, both in terms of accuracy and reliability of environment detection, and all necessary driving manoeuvres (lane keeping, lane changing, overtaking, stopping) as well as their planning – from a global route planning and optimization to the planning of the concrete trajectory for the next seconds/minutes – must be performed by the system; and this possibly even cooperatively with other road users. Control is only transferred to the human when the ODD is exited (e.g., when leaving the highway or at the depot). Additionally, AI can be used to predict typical road user behaviour and also to monitor adherence to the ODD.

Urban automated driving represents a fully automated vehicle. The ODD extends to municipalities and towns, as well as cities. Although speeds are lower in urban areas, there is no continuous structural separation between traffic directions and road users requiring special protection who are permitted within the shared traffic space. The consequence is a dramatic increase in the diversity of objects to be perceived as well as a temporal variability of the context due to the open context (e.g. through the introduction of new means of transport). The system thus requires advanced environment awareness regarding the objects (known and unknown) and their intentions (both typical and atypical behaviour). The planning of driving manoeuvres must be continuously adjusted, taking into account the future behaviour of other road users. Within the ODD, the system takes over all driving manoeuvres, especially in complex traffic routing with crossing lanes and dedicated lanes for other road users. Furthermore, monitoring is completely taken over by the system, and control is only handed over to a remote controller in emergencies (e.g. for recovery). In such a system, AI does not take on any new tasks; however, the demands on AI have greatly increased. As a special feature, the temporal variability of the context leads to the need for continuous system adaptation. To ensure the safety of the system, it is recommended to continuously monitor the operation and to create possibilities to exchange unknown scenarios (especially of environment perception) between manufacturers and vehicles.

From the perspective of applying AI methods in such increasingly complex and highly automated systems, minimizing potential dangerous encounters is achieved, for example, through an inner and an outer control loop. The inner loop contains all processes that can run automatically within the highly automated vehicle. The outer loop includes all processes that need to interact with the environment of the self-driving vehicle. In non-automated vehicles or at times when the vehicle is not controlled by automation, this is the task of the vehicle driver. In highly automated vehicles, intelligent AI components and systems are used for this purpose, which can distinguish and reliably interpret objects, processes, people, other vehicles, patterns such as light and dark, imprecision, and so on.

Modality aviation

Environment detection: In order to enable safe highly automated navigation and flight guidance, reliable environment detection is an elementary prerequisite. This usually requires not only the three-dimensional geometry of the environment, but also a semantic understanding of the environment due to the high safety requirements. This is especially true during take off and landing procedures and in ground-level flight. In these cases, depending on the aerial vehicle (AV), the requirements for semantic understanding of the environment are comparable to those of highly automated driving in road traffic. However, navigating in three dimensions results in a larger solution space. Especially in the development of drones and helicopters, not only the forward and downward field of view plays a role here, but the detection of a spherical 360° environment. This poses particular challenges for sensor technology, AI-based evaluation and associated computing resources in terms of the fields of view to be covered. In contrast, cruise speeds result in complementary requirements that must also be met by the interaction of sensors and AI. Here, the focus is on the area of “detect and avoid” (D&A). For D&A, there are significant differences in the required detection range and processing latencies, depending on the cruise speed, and the requirements for reliability of obstacle detection and classification remain very strict.

Trajectory planning: The use of AI for trajectory planning is being explored in several areas. In this context, deterministic AI methods are in principle capable of finding trajectories, but here too there are far-reaching requirements in terms of safety and reliability. In addition, environmental conditions such as weather and (up)wind as well as the associated effects of flight physics play an important role, making it difficult to validate such methods in practice. Furthermore, especially in local trajectory planning, algorithms work with data from environment detection, so algorithms have to deal with the corresponding uncertainties and take into account the three-dimensional fields of view that can be detected. For example, complex inflow conditions can lead to atypical flight situations or even high side slip (e.g., VTOL-G drones), where observability of the environment cannot always be adequately ensured.

Decision-making/flight planning: Certain decisions must be made even before departure. For example, the flight route must be determined, no-fly zones must be observed, and weather and other environmental influences must be taken into account. In principle, AI should be able to take over or at least support such functions in the future to enable a highly automated transport of people and goods. In order to be able to make such functions as well as other mission-relevant decisions during flight in a next step, taking into account external influences and other boundary conditions, further trustworthy AI methods are required that are capable of making meaningful higher-level and comprehensible decisions with high criticality.

Emergency landing: The behaviour of the AV in emergency situations is crucial for the safety aspects described above and for the assessment of risk. Since a failure, an uncontrolled landing or even a crash of the AV entails enormous consequences and it cannot always be ensured that predefined secured landing points are accessible or can be approached, the capability for an emergency landing represents an elementary safety aspect. In order to implement an emergency landing, functions from the areas of environment detection, trajectory planning, as well as decision-making must be combined. First, a safe landing site must be identified in potentially unknown terrain, and then a safe trajectory must be identified and approached. Many of the AI applications listed above play an important role in this, with many decisions being made based on environmental detection. Therefore, the implementation of a safe, redundant, and trustworthy environment detection is of particular interest for this modality. In this context, the technologies can also be applied to other use cases such as automated landing of drones, helicopters or aircraft at their destination or the identification of risks on the ground (reduction of ground risk by avoiding overflight).

General AI application fields: Other areas where AI can play a critical role include predictive maintenance and estimating battery conditions and remaining capacity. As described above, the failure of an AV due to faults in these AI systems usually results in a crash of the AV, which is one of the major differences from ground-based mobility and largely justifies the certification, safety, and redundancy requirements.

Package drones: A well-known use case is the delivery of packages by drone. Package delivery can be designed both in the form of a point-to-point mission (courier service) but also as a distribution of packages in the environment of a logistics hub. Likewise, the collection of shipments from the customer is also conceivable. Currently, these applications are being intensively researched, with the first companies taking their concepts to market entry. Especially with regard to the delivery of the packages, different concepts with different automation requirements are addressed. Drones could perform a parachute drop, as well as land in designated areas when delivering the package. But here, too, further degrees of automation are conceivable in the future. Equivalent to identifying safe emergency landing sites, drones can search for safe landing areas in unknown terrain surrounding the targeted address/coordinate to drop or pick up packages. Human interaction with the drone is also conceivable, with the human indicating the landing site or the drone responding to the human's gestures. Basically, automated package drones must be able to safely identify people and critical environments in all scenarios and reliably avoid any potential danger. This applies to both direct landings and package drops.

Automated air cabs: In the vision of the automated air cab, various manifestations are conceivable, from the point-to-point transport of passengers via hubs to the flexible use of air cabs as "robot taxis." The simplest case is regulated transportation from one hub to another, where take off and landing areas are controlled and the flight path is known and can potentially be secured. However, as soon as the application area differs from this, the demands on the flexibility of the AI functions for environment detection, trajectory planning and decision-making increase significantly. The requirements for coordination and communication also increase with increasing frequency in the lower airspace. Furthermore, the implementation of D&A functions becomes more relevant.

Modality shipping

In the near future, highly automated ships will increasingly be sailing the world's oceans. However, highly automated does not necessarily mean unmanned. It should be assumed that the degree of autonomy may change during a sea voyage. The range of possible use cases is wide and will vary greatly depending on the application. However, the number of crew members remaining on board will be reduced and the ship's bridge will not necessarily be manned during normal operations. The course of the journey is increasingly monitored and actively controlled from land infrastructure. Corresponding ships will travel long distances entirely without human intervention, making and executing nautical decisions autonomously within the operational design domain (ODD). The possibility of remote control will take on an important role and represents a functional level on the one hand and a fallback level on the other. An essential prerequisite is the reliable recording of the immediate vicinity of the ship in the near and far range in order to have an up-to-date picture of the situation available at all times. In addition to other vessels and their intentions, objects floating in the water and navigation signs must also be recognized safely under often changeable and difficult environmental conditions. In particular, information from computer vision systems and associated sensor technology will complete the situational picture and replace the eyes of the navigator. However, situational awareness is essential for safe navigation. This allows AI-based and continuous risk assessment, as well as dynamic and a collision avoidance rule (COLREG[93])-compliant adjustment of trajectories to minimize potentially dangerous vessel encounters. In the event of a sudden exit from the ODD or OOD operations, remote control can be quickly transferred to a remote operator. Since ships cannot always assume a safe state in the event of malfunctions or system failures, redundancies must be in place.

A detailed situational picture is a basic requirement for automated sailing in maritime shipping. In addition to position- and motion-specific data, semantic information must also be captured by the sensor system. Furthermore, collision avoidance regulations and environmental conditions such as weather and currents must be included in the situational picture. The AI-based functions are diverse and range from object detection to sensor data fusion and evaluation. Currently, there are neither suitable datasets nor established sensor combinations. There is a lack of specific provisions regarding data quality and its scope in order to develop appropriate models. As in the other modalities, reliable and meaningful verification and validation (V&V) methods are needed.

The exact route of the ship is usually determined before the start of the trip. Depending on the situation, however, there will always be minor deviations during the course of the voyage, for example in encounters with other ships. The determination of trajectories, some of which are complex, is currently rule-based, but in the future there will be more data-based approaches to better embed environmental influences and situational awareness into the decision-making process. A uniform model for the description of maritime traffic situations is needed, and/or a description language with corresponding interfaces. Critical scenarios must be identified to map relevant traffic scenarios, and simulations and virtual test sites are needed for testing. In-situ testing is only possible to a limited extent due to the diverse and cost-intensive framework conditions.

Modality railway

Non-system obstacles often cause unavoidable collisions, as the braking distance of the train drivers can be longer than the maximum geometric visibility range. The visual range of the train driver can be shorter than the maximum geometric visual range – it is proportional to the size of an obstacle. In addition, environmental conditions reduce the visibility of the driver. Dynamic obstacles can appear and disappear arbitrarily outside and inside the braking path. Even if a collision is unavoidable, emergency braking must be performed with additional whistling to reduce damage. It is never too late to carry out an emergency brake [338]. Damage reduction results from collision deceleration for dynamic obstacles, balance reduction, and hazard zone reduction at the point of collision. Even emergency braking triggered after the collision prevents a collided train from continuing its journey. However, the measure of emergency braking is problematic in the railroad sector. In the event of a false alarm, it cannot be cleared on all trains before they come to a standstill and is therefore associated with an economic loss. This state of affairs places high demands on the rates of false-negative and false-positive detections of an AI system used in the railway sector. Mathematically, the risk assessment even results in a much smaller tolerance for false-positive than for false-negative detections.

Most deaths in the EU railroad context are not accidents but result from trespassing with suicidal intent [339]. Trespassing accidents without suicidal intent are the largest category of fatalities relevant to risk assessment. In both cases, brakes must be applied, even if the reduction in damage within the braking path is small. The remaining accidents are considerably less frequent and can be greatly reduced by securing and reducing the number of level crossings. Reducing speed when colliding with heavier obstacles and avoiding further travel after a collision reduces the risk of derailment. Occupants of road rail vehicles survive collisions at lower speeds. Collisions with light non-human obstacles (e.g. birds, small branches and small land animals) are accepted without reaction.

[93] Convention on the international regulations for preventing collisions at sea.

In addition to obstacle detection, collision detection and condition monitoring are needed. Collision detection prevents further travel after collisions with undetected obstacles and enables the vehicle to restart after a false alarm. In GoA4, condition monitoring of the vehicle is a function of the driver and must be represented by automatic systems. Condition monitoring seamlessly transitions to predictive maintenance, using sensors and algorithms to monitor vehicles and infrastructure. At present, however, the maintenance intervals for safety-relevant components are fixed, i.e. independent of the observed condition. Since automated systems cannot replace the train driver completely, they must compensate for human advantages with capabilities such as long range obstacle detection (LROD). With LROD, a system should detect obstacles from a greater distance than a train driver can. Due to low accident frequency, high hurdles for access to the infrastructure and constraints for experiments, the collection of relevant perception data in the railroad sector is considerably more difficult. By applying condition-based and predictive maintenance, it will be possible to achieve flexibility from currently rigid inspection and maintenance deadlines anchored in the regulatory framework. However, this requires reliable statements about the quality and reliability of AI methods in order to be able to adapt the corresponding regulations and standards.

Challenges

The complexity of the functions in highly automated mobility systems – some of which are based on AI processes – and, in particular, the complexity and dynamics of the environment in which these systems must operate, inevitably mean that complete validation and complete proof of safety of all the system's behavioural possibilities in all conceivable scenarios under all possible environmental conditions cannot be approximately realized.

In the railway sector, an exemption from the Federal Ministry for Digital and Transport (BMDV) is required for automation levels from GoA3 on in accordance with § 45 I, § 3 I 1 of the Eisenbahn-Bau- und Betriebsordnung (EBO) (German Railway Construction and Operations Act). Approval of GoA3+ must be in accordance with the Common Safety Methods on Risk Assessment (CSM-RA) [340]. In simple terms, CSM-RA approval can be achieved following one of three possible paths – through a standard, according to harmonized design objectives, and by comparison with the "human" reference system. The first path is currently not suitable for AI-based obstacle detection systems due to the lack of standards. The existing standard, DIN EN 50657:2017 [89], for software on rail vehicles only covers conventional software. There have been several developments of obstacle detection systems worldwide using the conventional image processing methods [335]. One of the first such experiments was the KOMPASS project of the Federal Ministry of Education and Research in 2003 [341]. It is currently not possible to assess whether obstacle detection that can be used in operation is possible with conventional software. Therefore, the first path requires standards for AI systems.

In the automotive and the marine and aviation industries, as already indicated above, operational design domains (ODDs) and scenarios (see [342]) must be defined for each system within which the system may be used, and as part of the type approval process; the use of the system in these ODDs must be adequately tested and trustworthiness properties, such as functional safety in particular, must be demonstrated. For the latter, scenario-based testing approaches seem to provide sufficient assurance. On the other hand, the systems must be developed in such a way that (a) they continuously check at runtime whether they are actually still within the ODD, whether they recognize the current situation with sufficient accuracy, and whether errors occur in the implemented functionality; and that (b) in the event that the runtime test fails, they can fall back to a safe fallback level, i.e., to an operating mode with possibly limited functionality that at least allows a safe state (e.g., "stopping at the roadside") to be reached despite the unknown situation and leaving the ODD. Failed runtime tests that are not caused by "intentional" exiting of the ODD (e.g., intentional exiting of the ODD "highway" when the target exit is reached) ideally result in current system and environment data being reported back to the manufacturer or a central location, where it can be used to improve or further develop the system.

Overall, this means that both the development processes and the analysis and test procedures required for type approval or certification of such systems must be expanded in such a way that they allow the continuous (further) development of such systems, including update capability, associated runtime testing, and the appropriateness and functional safety of the selected fallback levels, and thus enable agile, continuous approval or certification of these systems (and their updates/further developments). For AI systems or systems with AI-based components, this results in particular in the challenge of proving the functional safety of these components with the accuracy required for type approval (and possibly other quality properties such as reproducibility, etc.) – this applies both to the type approval of the system, the certification of any necessary updates, the necessary runtime tests and the functional safety of AI functionalities implemented within the fallback levels. Some sub-challenges associated with this overarching method change are highlighted below:

As is necessary for the development of any mobility system, the first step for automated mobility systems is requirements elicitation and system (functionality) description. New challenges arise in the complexity and scope of the requirements, in which the system behaviour must now be described in relation to a significantly more complex environment and dependent on the behaviour of other transport users in this environment. Another innovation results from the necessity to specify the distribution of the driving task between human and system (handover times, handover modalities, possibly monitoring the attention of the user or the ability and willingness of the user to be able to take over the driving task). This continues with the need to specify the vehicle environment sufficiently well and accurately: This includes the description of the planned or allowed operating domain (ODD) of the future system, as well as a description of the relevant objects and artefacts that may occur in this operating domain – including the establishment of an ontology usable for object recognition It is to be expected that these descriptions cannot be complete, since a complete description of the artefacts occurring in reality is not possible either for the ODD description or for the relevant objects; however, a systematic process should be found, also across manufacturers, which, in addition to criteria and requirements for the quality and completeness of these specifications, also defines binding standard quantities for objects and artefacts. Since full testing of these systems is no longer possible due to the complexity of the environment (or ODD), the approach of scenario-based testing of such systems is currently the most promising approach and the one already required by various existing standards and those in preparation. An additional challenge here is the specification and description of these scenarios, first as a specification of the realized system functionality, but then in particular also as a basis for the tests and validation procedures to be performed. For the latter, the challenge is first of all to derive a sufficient number of test cases from the scenarios – in such a way that as large a test space as possible is covered within the ODD; this also requires, in particular, the combination and recombination of the behaviours of transport users described in a scenario with different environmental conditions such as weather, lighting conditions, road surface conditions, and much more. Furthermore, the “edge cases and corner cases” must also be identified from these relevant scenarios, i.e. those test cases that lie “at the edge of system performance” and in which the occurrence of errors is therefore most likely. Here, too, it is desirable to define a systematic process for scenario collection, for finding relevant scenarios specific to the application in question, together with the associated edge cases and corner cases. Due to the abundance of possible scenarios, processes that can perform these activities – in particular the collection of scenarios as well as the application-specific generation of test data from these scenarios – in a largely automated manner are to be preferred here; in this context, care must be taken, especially during scenario and test case generation, to ensure that inadequacies in the underlying data (such as bias or similar) are either detected and corrected or at least do not lead to corresponding properties in the AI functionality. The collection of these scenarios in manufacturer-independent databases could be advantageous, in order to avoid having to repeat the effort of scenario collection for each manufacturer, and also in order to be able to define uniform test criteria – i.e. uniform test scenarios per use case – as a minimum requirement.

Certain AI applications, especially safety-critical ones, also cannot be fully tested in the field or with real-world data. The critical events that must be used, for example, in a test of driving automation functions of higher autonomy levels are too different and can therefore be observed too infrequently. In addition, for obvious ethical reasons, testing by allowing critical situations (“child runs in front of car”) to occur is out of the question, even if they would not be evoked in the context with corresponding consequences. Therefore, it is necessary to use synthetic data, which are generated from a simulation (“digital twin”). In the simulation, critical scenarios can be generated specifically, at least those that are known or occur in the field at least with a certain frequency. In addition, the simulation offers the possibility to set all relevant parameters up to the limits of what can be physically expected, so that critical scenarios can be generated that were previously completely unobserved but are nevertheless possible. Compared to field testing, this approach also has the advantage that events are reproducible, which greatly facilitates the analysis of the results obtained.

Regardless of how high the level of automation of a system is, suitable fallback levels must be in place to allow safe operation or safe stopping of the system even in the event that the system can no longer perform its task in a functionally safe manner. This case can occur in systems such as the highway chauffeur, for example, when the system detects that it can no longer recognize its surroundings correctly due to weather conditions, for example, so it issues a handover request to the driver, but the driver does not take over the driving task. But even in fully automated vehicles (SAE Level 5), the failure of subsystems – e.g. stone impact on the camera used for environment perception – or other factors such as the unexpected departure from the ODD may mean that the driving task can no longer be fully performed by the system. Currently, safe fallback levels typically consist of performing an MRM (minimum risk manoeuvre; e.g., pull over to the right side of the road and stop). In the future, more complex fallback levels (e.g., continuation of the driving task with possibly greatly reduced speed) may be possible. In addition to the challenge of continuous self-monitoring, which the system must perform to determine whether it can still fulfil the driving task, the definition of suitable fallback levels – which should ideally be dependent on the specific error that has occurred – and the safety verification of these fallback levels are currently unsolved problems.

Special provisions must also be made for the handover of the driving task. Here, the handover to the system is typically simple and is often done by the driver manually switching on the appropriate automation function. Typically, the system tests whether it is within its operating range and is functional, takes over the task, and confirms this takeover. The transfer of the driving task to the driver is more complex. Normally, windows of time are given within which the driver must accept a request to take over (otherwise an MRM is triggered, see above). The challenge is then to monitor during operation that the users are sufficiently alert to allow acceptance within this time window if a handover request becomes necessary. Furthermore, within this time window (i.e., after a situation has occurred that has led to the request for the driver to take over), the system must still be able to safely perform the driving task – often even the minimum risk manoeuvre executed afterwards. Depending on the specific situation and the concrete conditions that led to the request for takeover, the functional safety of the system can often only be ensured with great effort during these periods. Finally, for the handover of the driving task to the driver, it must also be examined whether a slow, transient, and partially assisted handover to the driver – despite the presence of a situation triggering the handover request – would not increase the overall safety of the system.

To meet this abundance of challenges, the establishment of a uniform, possibly even cross-manufacturer infrastructure to support the continuous further development of the highly automated mobility systems described above makes sense.

4.6.3 Standardization needs

Need 06-01: Record the state of trustworthiness-by-design and of testability
As a basis for operationalizing the planned EU AI Act (currently in the 2nd revised draft version), comprehensive guarantees with regard to the trustworthiness of AI technology in the field of mobility represent a major challenge, and necessary legal and organizational frameworks and technical methods and tools are currently not sufficiently available for practical use. The parameter space to be considered is highly dimensional, including: a) the life cycle phases and embedding of the AI system, b) the different trustworthiness aspects (TW aspects: safety, security, robustness, transparency, fairness, explainability, etc.), c) the different AI models and learning methods, and d) the different use cases (domains, modalities and functionalities) in the mobility domain.

The state of development and testability of AI systems should be systematically recorded for relevant applications in the field of mobility with regard to the parameter space mentioned above and tracked over time in order to meaningfully prioritize further research and development (R&D) work, especially in the areas of "X-by-Design", testability and verifiability of system and component (trustworthiness) properties such as IT security, reliability, explainability and introspectability, as well as measures for safeguarding.

The standards and specifications to be established should specifically cover:
→ the establishment of a method that allows objective comparison of trustworthiness-by-design development and testability with respect to different applications and over time,
→ concrete, practical evaluation criteria for the entire relevant parameter space,
→ an explanation of the method with concrete examples.
→ As far as possible, this should be based on existing standards and specifications, e.g. ISO 21448:2022 [90].

Due to the high leverage effect (avoidance of duplication of work, use of synergies and focus on essential work), it is imperative that policy-makers provide the necessary resources for this need.

Need 06-02: Development and practical implementation of lacking technical, legal and organizational fundamentals

Due to the high level of complexity (see Need 06-01), many technical, legal and organizational fundamentals for trustworthy-by-design development, testing and safeguarding in operation are either not feasible or are not sufficiently feasible and available. However, these are a prerequisite for sufficient guarantees regarding the trustworthiness of AI systems in the field of mobility. The technical, organizational, and legal fundamentals regarding the trustworthiness of AI systems in the context of mobility (cf. Need 06-01), which have been lacking or cannot be implemented in a sufficiently practical manner to date, should therefore be systematically developed and implemented in a practical manner. This includes, in particular, suitable metrics (key trustworthiness indicators) as well as the definition of minimum qualities based on these metrics (e.g., "acceptable residual risk"), vulnerabilities, interpretation methods, safeguards, responsibilities and their respective dependence on the boundary conditions (ODD).

The standards and specifications to be established should specifically cover:

- → Technical, organizational and legal fundamentals for all practical combinations of life cycle phase, trustworthiness aspect, use case / functionality and AI technology.
- → Boundary conditions under which these fundamentals are valid, and practical guidance on how to adjust the boundary conditions to increase trustworthiness.
- → This should build on relevant existing cross-sector (especially DIN SPEC 13266:2020 [98]) and sector-specific standards and specifications (e.g. ISO 26262 series [455] for road; DIN EN 50657:2017 [89], DIN VDE V 0831-101:2022 [344], DIN VDE V 0831-103:2020 [343] and DIN EN 62267:2010 [332] for rail) and sector-specific standards and specifications should be extended accordingly (e.g. ISO 21448:2022 [90] extended to cover rail and other mobility sectors).

Without a speedy development of these fundamentals, a timely operationalization of the AI Act ([4], currently 2nd revised draft) will not succeed, and therefore it is imperative that policy-makers allocate the necessary resources for this need.

Need 06-03: Generalized requirements catalogue that can be easily adapted to specific domains and use cases

Due to the complexity of the use of AI technology in the field of mobility, specific requirements cannot be formulated in advance for all combinations of life cycle phases, trustworthiness aspects, AI technologies and applications or functionalities.

A generalized modular requirements catalogue with regard to technical, organizational and legal aspects is to be developed together with practical instructions and concrete examples for adaptation to any use cases in the field of mobility. Specific adaptations to concrete applications and functionalities as well as to boundary conditions for these applications (e.g. safety-critical applications) should be possible with as little effort as possible.

The standards and specifications to be established should specifically cover:

- → a comprehensive modular and application-agnostic catalogue of requirements with regard to technical, organizational and legal aspects,
- → detailed instructions and practical examples for the application-specific adaptation of the requirements catalogue, where „application-specific" includes both the concrete requirements and characteristics of the application and the application's deployment environment (e.g., safety criticality),
- → guidance on the favourable design of framework conditions to increase trustworthiness on the one hand and to reduce development and testing efforts on the other,
- → architectures and architectural models to reduce the propagation of uncertainties and increase accessibility for detection methods,
- → methods for the introspection and proof of safety and reliability of AI,
- → application-specific risk acceptance criteria.

Need 06-04: Continuous (further) development and validation in operation

As outlined in Chapter 4.6.2, both the development processes for highly automated mobility systems and the analysis and test procedures required for type approval or certification of such systems must be expanded to allow the continuous (further) development of such systems, including update capability based on data collected in the field, associated runtime testing, and the adequacy and functional safety of the selected fallback levels. For AI systems or systems with AI-based components, this results, in particular, in the challenge of proving the functional safety of these components in a precision and scope necessary for type approval – this applies both to the type approval of the system, the certification of any necessary updates, the necessary runtime tests and the functional safety of AI functionalities implemented within the fallback levels. Overall, these processes and procedures must thus allow dynamic, continuous (re)certification or type approval in the sense of continuous system development.

The standards to be established and the existing standards must support these development processes and define requirements for type approval, supporting in particular the update capability and the backup concept implemented via runtime checks and fallback levels. In addition to the special consideration of AI components in the above sense, it should also be possible to catalogue findings in the form of critical scenarios across manufacturers. On the one hand, this serves to continuously improve the systems and, on the other, to sharpen the security requirements (e.g., for domain shifts).

The specifications and standards to be established should therefore specifically cover:

→ systematic identification processes for critical scenarios,
→ multi-manufacturer interfaces, exchange processes and specifications for an ecosystem with independent bodies (especially for scenario catalogues),
→ specifications for monitoring, testing, safeguarding and certifying systems with AI components within a continuous development and update process,
→ best practices for mitigating AI system malfunctions in mobility,
→ guidelines / best practices for safe/trustworthy-by-design development for relevant use cases (see use cases column) or ideally generalized recommendations with concrete advice for application-specific adaptation,
→ specifications for safe fallback levels including a takeover of control by humans,
→ recommendations for action to define responsibilities in the development, testing and practical use of AI technology in mobility.

Need 06-05: Analysis, simulation and test methods as well as test infrastructure

The complexity of applying AI technology to mobility requires a) interdisciplinary knowledge, b) standardized methods and tools, c) large amounts of quality-assured data, often subject to limitations on use, e.g., regarding privacy, and d) large computational resources for simulations, training, and testing. This requires the use of simulative methods, and a testing infrastructure whose requirements can only be met by very few large corporations or state actors. This requires close cooperation with many partners, including information exchange, joint projects, and shared use of data and computing resources (see also [345]). In order to enable such cooperation and also to establish comparability of automated mobility systems and the AI components used, especially with regard to trustworthiness and safety and their verification, not only the minimum requirements for the systems must be standardized. Rather, it is also necessary to define appropriate methods to support the development of the systems and verification of their properties. In this context, simulative methods provide a cost-effective and non-threatening way to support the development and verification of AI components and systems with AI components for mobility solutions. Without standardizing these and making them verifiable through quality criteria, comparability of results cannot be achieved. Therefore, also in many application domains, the development and provision of a collaborative (virtual or physical) test infrastructure is useful to easily enable a close interdisciplinary and international exchange of information, the sharing of data and computational resources for development and testing in simulation and the physical world, and the exchange of methods and tools. In terms of functional safety, the quality of the simulation procedures is of particular importance; here it must be ensured that the simulation has a sufficiently high correspondence with reality in order to obtain reliable statements for the type approval and certification of these systems. Currently, neither methods nor chains of argumentation exist that guarantee this correspondence to a sufficiently high degree.

The standards and specifications to be established should specifically cover:

→ virtual simulation and test methods, test environments and their quality,
→ methods for verification and validation (especially extension of the „Sotif standard“ ISO 21448:2022 [90] to other domains such as railroads),
→ guidelines for AI certification, and development and test methods,
→ standardized terminology for efficient communication,
→ standardized interfaces for the exchange of data, models and simulations,
→ standardized procedures for the shared data management, development, and testing of AI systems.

Need 06-06: Scenarios, datasets, interoperability, interfaces, data exchange, data quality, digital twins
An exchange and comparability of AI components and their trustworthy and safe use in automated systems requires standardized interfaces and minimum requirements; the concrete design of these interfaces and minimum requirements to be standardized is, at least in part, the subject of active research. These interfaces should primarily define the use within test environments and, together with the requirements, also the operation within the systems used for test execution. Here it is necessary to define the minimum range of use for different applications in a uniform way. Due to the complexity in the mobility domain, scenarios are considered as a means of describing and structuring the intended operating domain (ODD) at the system level. To enable comparability of of the fulfilment of requirements, the criteria for critical scenarios and (field) datasets and scenario catalogues used in testing the systems must also be standardized, and to achieve interoperability and enable data exchange, the exchange formats for field data, scenarios, and datasets must continue to be standardized. For standardized component tests, especially within perceptual testing, standardized sensor configurations, which have to be provided by a test environment, are still necessary. The use of digital twins and quality requirements, especially for the generation of synthetic data, should also be standardized for the cost-effective addition of test datasets. Where necessary, appropriate research must accompany or precede the standardization work.

The standards and specifications to be established should specifically cover:

→ a uniform description of the ODD, scenarios and, if applicable, interfaces for different systems,
→ minimum requirements, specifications, and supporting exchange formats for scenarios and datasets,
→ criteria for labelling data and scenarios and for covering the ODD,
→ criteria for critical scenarios (especially related to safety, but application-related also to all trustworthiness aspects),
→ standard datasets (including edge cases and corner cases), standard sensor configurations, and dataset quality requirements,
→ best practices for the generation, quality assurance and use of synthetic data,
→ requirements for the traceability of data used, e.g., to prevent backdoors from being introduced (poisoning backdoor attacks).

Need 06-07: Testing with synthetic data
With regard to standardization, the question arises as to how to ensure the validity of the synthetic data used for testing. The question of the data used for training can initially be considered to be secondary, because if the testing is rigorous, comprehensive, and valid, it will implicitly reveal the quality of the training data. The synthetic test data must have a sufficiently slight difference from the data found in the field. However, this difference does not relate to the subjective impression of "realness" that a human observer has. Rather, objective and task-specific measures must be used. Even with a camera-based perceptual component, it is not a foregone conclusion that a synthetic image must appear as real as possible, or that this alone would be sufficient to ensure validity. Furthermore, in a behavioural component (driving strategy and trajectory planning), synthetic as well as real data are available in a much more abstract way, e.g., as position data of aligned rectangles that change over time. The task of standardization is therefore to determine for all relevant components (use cases) under which conditions synthetic data correspond to real data to a sufficient degree. The validity of the synthetic data must be continuously checked, since an extension of the underlying scenarios or an adjustment to the changing actual conditions (road topography, etc.) may cause differences that did not occur before. The validity test must also take into account the fact that no one-to-one comparison of all scenarios can take place. Rather, it must be ensured that the extrapolation (scenarios that are exclusively synthetic) is valid. In addition, it is necessary to determine when a test has sufficient coverage of critical scenarios. In addition to standardizing the approach to generating synthetic data for testing automation functions with AI components, it is therefore conceivable that regularly updated scenario catalogues will be held by independent bodies.

The Working Group Mobility ranked the identified needs according to the urgency of their implementation. Figure 42 shows the urgency of implementation, categorized according to the target groups of standardization, research and policy.

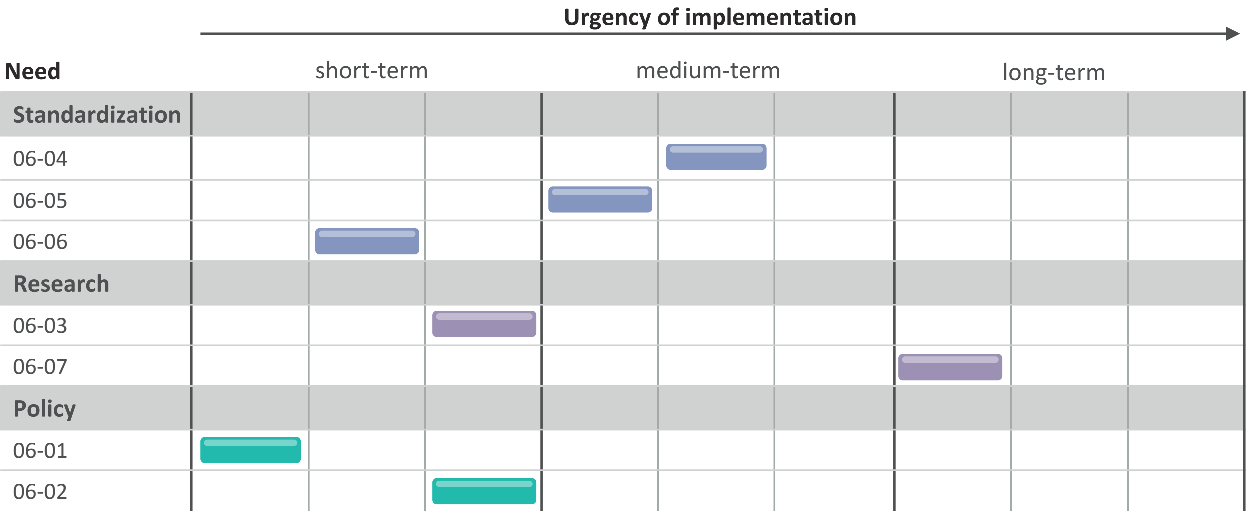


Figure 42: Prioritization of needs for the key topic mobility (Source: Working Group Mobility)

4.7 Medicine

The use of AI to improve medical care is one of the application areas that the European Union (EU) sees as a key application field with great potential [7], [346]. The use of AI in medicine for the purpose of diagnosis, screening, therapy (recommendation), monitoring, triage, and prognosis of diseases occurs both in lightly regulated areas to optimize the organization of healthcare facilities, the healthcare system as a whole, or general health apps, as well as in the heavily regulated areas of medical devices. The issues presented here apply analogously on an ongoing basis to in vitro diagnostic.

AI-powered algorithms are capable of analyzing large amounts of multimodal data, and in doing so, are able to identify patterns within a relatively short period of time that humans would have limited ability to do. AI systems can already outperform human experts in individual medical tasks (e.g. skin cancer screening; [347]).
Strict safety requirements must be met before a new product can be used on humans, particularly in the case of medical devices. This makes the development, implementation, and conformity assessment process required for market access of AI-based medical devices a complex process with multiple regulatory, ethical, technical, and clinical requirements In the meantime, a number of medical AI applications have successfully undergone such conformity assessment procedures and have already been successfully placed on the market (see, for example [348], for AI medical devices in the EU and USA, and all AI products cleared by the Food and Drug Administration (FDA) [349]). For the highly data-driven approaches of AI- or ML-based systems, – and compared to non-AI-based systems, there are specific aspects that need to be considered in new or extended ways in order to successfully pass the conformity assessment process: Examples include quality of data and real-time decisions, reliability of results, complexity of models, effective integration with existing clinical workflows and IT systems.

For the field of AI-based medical devices, generally applicable standards and specifications need to be developed, which for the most part do not currently exist at the national, European or international level. In this context, it is difficult to take a generic view that integrates all facets of the application of AI in the field of medicine. In the following, three use cases from the fields of medical imaging, dentistry and intensive care medicine will be discussed in order to derive needs for action for the development of suitable procedures and standards.

4.7.1 Status quo

The conformity assessment of medical devices is centrally regulated in the EU in the Medical Device Regulation (MDR, [350], as of 2021). For the requirements described there, there are already a number of standards that cover central aspects such as quality management [381], risk management ([351], [352]), software life cycle ([353], [354]) or usability ([355], [357]) that have been established in the medical technology industry for some time. These standards implement general requirements for medical devices, but do not include specific requirements for AI-based systems. In parallel, there are horizontal, i.e., cross-industry, sets of rules for implementing AI-specific requirements, such as the IEEE 7000 series of standards (2021) [10], [11], [12], [13] or those currently being developed in ISO/IEC JTC1/SC42. However, these do not consider specific requirements for medical devices and can only to a limited extent fill existing gaps regarding the increased requirements in medicine.

To nevertheless obtain reliable procedures for the implementation of AI-based medical devices and their conformity assessment, the Association of Notified Bodies for Medical Devices in Germany (IG-NB), for example, has published a guideline “Artificial Intelligence in Medical Devices” [358], which systematically records core requirements for AI-based medical devices and thus provides assistance for the conformity assessment process. Many notified bodies rely on this questionnaire as a key reference when reviewing AI systems. Currently, this assumes that an AI-based system always has a frozen state when it is evaluated. Further learning after commissioning at the customer’s site would consequently require a new conformity assessment as soon as substantial changes are made to the AI system. There are currently no normative or regulatory provisions, which defines substantial change in a practical fashion. Likewise in the U.S., there are also no specific rules governing the regulation of AI-based medical devices. The U.S.-FDA made a proposal for regulating AI-based medical devices in April 2019 with [139], but as in Europe, it has not yet been translated into concrete guidance documents. Nevertheless, this proposal does include AI systems that continue to learn during operation. Based on a fixed state, it is already possible to bring AI-based medical devices to market. This can be seen, for example, in the list of over 300 products now cleared in the USA alone (see [349]). There are also systems already on the market in Europe and Germany.

The planned AI Act will formulate requirements in the EU that address AI-specific aspects in a legally binding manner in the future. The numerous feedbacks from relevant market participants have made it clear that there is a need for further harmonization with existing regulatory requirements such as the Medical Device Regulation (MDR). If, for example, the contradictions between the MDR and the planned AI Act, are not resolved, additional costs must be expected or market access will even be denied, since the requirements of the MDR and the draft AI Act must be implemented simultaneously in the conformity procedure (see also Chapter 1.4 and Annex 13.1, Chapter "Exemplary presentation using the example of medical devices").

In principle, it should be noted that AI-based medical devices have some special features that do not come into play in the same way in other areas of application and therefore must also be considered separately in standardization. These include the following core aspects:

→ HIGHLY PERSONAL DATA: Medical data is usually highly personal and often extends into sensitive areas. For Europe and especially Germany as an important location for medical technology, it should be noted that due to existing data protection rules and additional data protection laws at the federal and state levels access to data is more heavily regulated compared to countries such as the USA or China. On the other hand, the MDR requires comprehensive data for demonstrating safety. In the meantime, the EU has also taken up this issue, for example, by drafting a European Health Data Space (EHDS [359]) to enable better access to medical data. Since the EHDS is currently still at the planning stage, some points remain unresolved. Among other things, this concerns the question of how differentiated access to health data can be ensured in the future for those directly affected and other stakeholders in accordance with the GDPR. This includes to what extent and under what conditions a company can be granted access to medical data for the purpose of developing commercial products. Very important and complex sub-issues are the use of anonymized vs. pseudonymized medical data and the problem of re-identification of personal information for certain types of data (especially image data, e.g., cranial imaging, as well as for very individual parameters, e.g., persons diagnosed with a rare disease in a certain institution) when anonymization is actually present.

→ RISK-BENEFIT CONSIDERATION: Before it is placed on the market, the decisive target value for a medical device is always a clinical evaluation of the fulfilment of the specified performance requirements, safety and patient benefit. Typical evaluation criteria used in other fields and still relied upon in many medically oriented AI publications generally cannot do this. For example, in a diagnostic test, minimizing overlooked diseases (false negatives with sometimes serious consequences) usually leads to an increase in too many misdiagnosed diseases (false positives), which can also cause harm such as patient uncertainty, unnecessary interventions, and so on. It is necessary to find a balance between these opposing effects, to include the respective effect of the different types of errors in an AI procedure, and ultimately to make an evaluation in terms of clinical success and to optimize the performance of the overall system. The planned AI Act, on the other hand, calls for individual risks to be reduced first and foremost. Approaches that map overall risk or the risk-benefit ratio are not present in the draft AI Act in the form that they should be implemented specifically for AI-based systems.

→ LIMITED AVAILABILITY/QUANTITY AND HIGH COMPLEXITY OF TRAINING DATA: High quality training datasets are critical to the performance of an AI system in its anticipated clinical setting. AI systems may not function properly if, for example, they are deployed in different populations or in a different context (e.g., different hospital) where they may be confronted with different data and circumstances than those with which they were trained. In addition, for some areas, such as surgery, data acquisition, e.g., via clinical trials, can be difficult and only a few cases can be included.
Thus, for certain types of treatments, only a very limited number of high-quality datasets are available, as these must come from dedicated studies in real-world application environments. In addition, individual factors and multiple aspects of the treatment environment often contribute to the success of treatment. When taking such variations into account, care must be taken to ensure sufficient statistical reliability in all relevant areas (e.g., with respect to patient populations, indications, but also different physician approaches and different hospital environments). In order to realize the added value of AI-based methods, especially in a more individually oriented treatment up to a personalized medicine, in addition to new requirements for the study design beyond classical-, statistical proofs, ways have to be found to make datasets available, e.g. via the generation of synthetic data or methods such as federated learning (AI models learn from decentralized training data sets, the data remain e.g. in the respective hospital). Clear guidelines in

particular on how to implement quality control are still lacking.

- → FORMALIZATION OF PARAMETERS FOR RISK QUANTIFICATION: In the case of AI systems in the medical field, the formalization and quantification of risk criteria is naturally subject to particularly high demands. This requires a conclusive classification of the risks, which is often difficult to achieve with new medical devices as long as the products have not yet been used in regular operation. However, in the case of medical devices, there is a requirement that sufficient clinical data (i.e., data from a real-world application) must be available or provided via clinical trials before the product can be launched on the market. For this reason, existing regulations (in particular DIN EN ISO 14971:2022 [351]) allow a gradual approach in the form of a semi-quantitative assessment of risks for pragmatic reasons.
- → VARYING DEGREES OF AUTONOMY AND HUMAN SUPERVISION REQUIREMENTS: It should be noted that the purpose of medical devices and the associated risk potential can vary greatly depending on the level of autonomy of an AI system, from a purely supportive to a largely autonomous system. Most AI-based systems currently under development or already in use are in the field of diagnostics or radiology (e.g. mammography screening, diagnostics of eye diseases or skin cancer) [348]. For example, in diagnostic applications, a human observer could always be used as an additional control instance before a final decision is made (human-in-the-loop). For other systems, e.g., an alarm system in an intensive care setting or a ventilation system that functions in an automated manner, human control would be largely eliminated at the highest level of autonomy and the AI would function as a closed-loop system. Such aspects would have to be systematically included in the risk assessment. To that end, the proposed AI Act also includes a requirement to integrate human oversight into products so humans can intervene in the operation of the system at any time However, it does not describe what such oversight can or must include. In addition, there is a lack of specifications as to what level of explainability AI systems must achieve in order to ensure sufficient security.

4.7.2 Requirements and challenges

Basic requirements regarding the implementation of conformity assessment for AI-based systems – according to the current state of standardization and legislation.

Medical devices, i.e. instruments, devices, software or similar with a dedicated medical purpose, are subject to the MDR in the EU and thus have to fulfil a wide range of requirements. AI-based applications in medical devices mostly fall into the software category and are to be classified (according to MDR, Annex VIII, Rule 11) as IIa for lower potential harm or as IIb up to risk class III for higher potential harm. In these cases, the products must be subjected to a conformity assessment procedure involving a notified body in accordance with the MDR. As a result, they also fulfil the criterion in the planned AI Act that leads to a classification in the class of high-risk products within the meaning of the AI Act draft (see Art. 6 and Annex II there). A large number of AI applications in the field of medicine will thus be subject to additional requirements of the planned AI Act in the future, in addition to the existing requirements of the MDR. Challenges that could arise from this in the future from an organizational point of view are presented in Chapter 1.4 and specifically in Annex 13.1 (clause “Exemplary presentation using the example of medical devices”). The focus of this chapter is on the basic requirements that an AI-based medical device must fulfil due to the existing regulations (especially with regard to the MDR). As described above, neither the MDR nor the associated standards contain specific requirements for AI-based systems. As a result, manufacturers currently have to make do with using unofficial guidelines such as the IG-NB questionnaire [358] to demonstrate the compliance of an AI-based medical device

To demonstrate the performance and safety of the product for the given intended use, corresponding requirements must be implemented on both the technical and clinical sides. On the clinical side, this includes a comparison with solutions that have already been put into operation and tested. In addition to the technical comparison of performance, this includes a positive evaluation in terms of the risk-benefit ratio. To this end, it must be defined to what extent additional risks (in relation to classical methods or the state of the art) are permissible and acceptable, and to what extent these risks can at least be offset by a corresponding clinical benefit.

The interaction between the users and the system must also be taken into account. This can prove to be complex due to interdependencies, especially in the case of AI-based systems. This is especially true , on the one hand, the user relies more on the reliability of the results and, on the other hand, the system adapts its behaviour to the respective application environment. Significant shifts in risk assessment can occur if, for example, users rely on specific diagnoses (see imaging use case) or alarms (see intensive care use case) determined by the AI system. Even if the results achieve better accuracy, relying on the results of the AI system can increase the risk-potential. Therefore, a distinction must be made between, for example, technical performance criteria (e.g., detection rates of critical situations) and clinical parameters (e.g., harm to patients due to incorrectly overlooked diagnoses or critical situations).

On the technical side, a number of criteria must be implemented to ensure the fundamental safety and security of the AI system. This includes determining the performance of the model based on technical criteria, providing/availability of suitable data for training, testing (validation, testing) of the AI-based model, ensuring the correctness and robustness (fault tolerance) of metrics on which the AI-based decisions are based, and aspects of software architecture such as. e.g., integration of 3rd party components for model training (i.e., third-party components that are integrated into the AI-based system), and cybersecurity (e.g., specific requirements for AI systems with respect to protecting the systems from tampering). Moreover, in terms of transparency, disclosure an evaluation of the basic clinical model on which the AI system is built (which individual parameters have which influence on the clinical decision) is an essential factor for the acceptance and implementation of conformity assessment for such AI systems.

It is important to note how the established criteria of AI interact with clinical benefit. Classical error criteria for training AI models such as “accuracy”, specificity or recall alone cannot directly classify how good the quality of the models is in a clinical context. For example, an overlooked alarm (false negative) can have considerably different effects than an incorrectly triggered and thus unnecessary alarm (false positive). In addition, how reliably clinical staff react to the alarms or diagnoses and how well they can classify them and understand their causes is crucial.

Individual requirements regarding the technical and clinical evaluation of the AI-based system

Describing and selecting the model

In addition to the selection of performance criteria, conditions must be attached to the model used in the medical context. Thus, within the framework of conformity assessment

- → the chosen approach needs to be compared with the state of the art in technology and medicine, as well as the use-case specific established “gold standard”,
- → robustness, fairness, and reproducibility need to be demonstrated,
- → the remaining uncertainty of the prediction needs to be adequately stated, and
- → the transparency of the prediction result needs to be promoted (keyword „explainable AI“).

The chosen approach of the medical device must be supported by extensive research on the state of the art and established methods in the medical context. Here, it is advised to use the metrics listed under performance criteria to compare the models.

In the remaining uncertainty investigation, limits to predictability should be specifically sought out and communicated. Here, marginal medical cases have to be researched and tested with the medical device at hand, and the quality of the prediction with normal findings has to be analyzed. For proof, a statistical examination of various criteria (random variables of the population) and the specification of confidence intervals is helpful. For example, robustness and fairness (in the sense of an existing bias) of a prediction result can also be described quantitatively. The robustness investigation also includes generalizability analysis, i.e., the applicability of the AI solution to data from devices of different manufacturers or in different application environments. This is to be carried out in the form of a check based on independent data, if necessary using cross validation, if only a small amount of data is available.

In medical applications, the acceptance of an AI medical product as a black box is only acceptable if particularly strict requirements are met. A simple traceability of results is often limited with AI methods and especially with neural networks. However, for specific predictions and decision-making, a rationale must be presented that meets the requirements for a medical device: A clinical user must be able to review the reasonableness of predictions and decisions and, if necessary, to correct htem. “Explainability” also includes the visualization of prediction with their reliability values, based on transparently presented fundamental rules/characteristics of the evaluation that the AI uses for such decisions. In the area of visualization during the application, there is also a strong focus on usability, an intuitive user-friendly or graphical interface. However, such approaches in turn require quality assurance, which is currently only in its infancy in this field.

Performance criteria

To assess performance in the context of technical validation of AI models that make a certain prediction as subcomponents of an AI system, an unknown representative dataset, also called a test dataset, is used to compare the model predictions with the annotations (set as the "gold standard" by human experts). For this purpose, classical performance criteria such as accuracy, specificity/sensitivity, precision/recall, F1 score, receiver operator characteristic (ROC) curves or area-under-the-curve (AUC) values are usually used. In addition, the associated detection probabilities (probability of detection – POD – or probability of classification – POC) can be included. This addresses, for example, how likely it is that a tumour of a certain size will be detected (POD) or that this tumour will be correctly classified as benign or malignant (POC). The criteria mentioned here mainly refer to tasks from the field of classification and supervised learning. Appropriately aligned metrics should be used for other tasks. This concerns quantitative estimates from the field of regression, but also more complex scenarios with dynamic aspects, such as those involved in optimized therapy planning (see, for example, use case intensive care medicine). For dynamic and time-critical use cases, in addition to performance, time must be taken into account. In some cases, the performance criteria can be supplemented by weightings to take account of specific risk factors.

Data management

The provision of the data basis for training and testing the AI model must meet several key quality criteria. This includes, for example, the aforementioned independence of training, validation and test data. Each of these groups must also provide representative coverage of test data relevant to the intended application. This includes consideration of different settings (e.g., different equipment, different qualifications of the care team and approaches to treatment processes, different infrastructure in hospitals, physicians' offices) and an appropriate range and representativeness of the patient population (e.g., age, gender, ethnicity). This requirement for AI training, validation, and test data also applies with respect to ethical issues such as fairness and inclusivity toward diverse groups or non-discrimination with respect to minorities. One of the difficulties here is to treat the different groups equally to achieve the best possible individual treatment, while at the same time being able to exploit the specific potential for optimization in the individual groups as extensively as possible. To enable a representative mapping of the patient groups defined by the purpose, a common understanding should be established as to which demographic variables have a significant impact on the clinical workflow for the specific purpose and consequently need to be represented accordingly in the training, validation and test data.

The overall provision of data must be at a very high level of quality, although standardized methodologies and tools for assessing data repositories have been lacking up to now. In supervised learning, for example, it is important to ensure that labelling of data is done by suitably qualified personnel. Several experts may have to annotate the data independently of each other to avoid a possible bias. In addition, clear process provisions must be made to perform labelling at an appropriately high level, even when supplementing data.

AI-specific questions of a risk analysis

DIN EN ISO 14971:2022 [351] requires medical device manufacturers to have a risk management process in place to ensure that risks from medical devices are identified, assessed and controlled, also, that this is always acceptable in relation to the benefits. This also applies to risks that arise in connection with AI-supported medical devices. In general, the risk analysis should also consider the following points:

- → Risk-benefit consideration between using AI or using classic AI-free methods (hard-coded decision trees).
- → The understandability and clinical evaluability of the prediction result by the prediction model (thus the result finding process) must be ensured to an appropriate degree. This applies not only to the inspection by a notified body during the conformity assessment procedure, but also in the event of notification during operation, including appropriate intervention options.
- → Measures need to be put in place to collect log data and vitality/operating information from the AI system that will allow an assessment of the functionality of the AI model to be made and, if necessary, identify malfunctions.
- → Quantity and quality of data available for training, validation and testing:
  - It is to be verified whether there is a sufficient amount of data usable for training, validation, and testing of the AI model is to be verified. If necessary, the data collection must be enriched by synthetic data.

- Documented evidence must be provided that the data is as free of bias as possible. If only a limited cohort of data is available, it has to be checked, whether this has an impact on the intended use of the product, and if the targeted patient population may need to be adjusted/restricted.

→ For the future implementation of continual or incremental learning AI systems:
  - The risk-benefit consideration for using an open versus a non-continual learning AI model needs to be done.
  - Risks specific to continual learning systems are identified and mitigation measures are implemented.
  - The system can be reset to a known training state.

Clinical evaluation

The clinical evaluation of an AI-based medical device must ensure that, when used in a complex clinical setting, the system is safe, performant, useful, does not cause unanticipated harm, and always provides the professional user with sufficient intervention options as regards the selection or decision criteria. Ethical aspects must also be taken into account. Clinical evaluation must occur at all phases along the life cycle of the AI-based medical device. At the beginning of the development phase, the focus should be on a review of the intended purpose as well as the state of medical practice in the application field of the AI-based medical device. This includes understanding the medical problem the AI application is trying to solve and whether the application is appropriate for that purpose. In addition, a description of the intended clinical benefits versus established methods, the potential risks and harms that could be caused by the AI, and good documentation of interoperability with both the user and, for example, the IT system, including a review of the "user experience" to include safety-related issues, is critical. During further development of the AI model underlying the medical device, it is necessary to accurately list the model test data used and compare model performance to the current gold standard. It should be noted that for some AI applications, a corresponding use-case-specific standard may still need to be defined.

Clinical evaluation also includes the generation of clinical data to validate performance, safety, and patient benefit. The performance of an AI system may be optimal under test conditions, but when used in "real life" it may no longer meet the intended benefit due to various human and technical influencing factors. Therefore, clinical data generation should be integrated as a necessary tool to evaluate AI technologies before and after their implementation as an important factor in development. This is also enshrined in the MDR, which very clearly calls for the availability of clinical data at an appropriate level.

When conducting clinical trials, it is important to map the effect of the AI intervention across the treatment pathway.

Clinical trials provide the necessary evidence for the efficacy and safety of a medical device/AI system. There are several types of trials that differ in scope and procedure, each having certain advantages and disadvantages; for example, randomized clinical trials (randomized controlled trials (RCTs) or cohort trials) and retrospective case-control trials are among the most important. The selection of an appropriate trial type also depends on the explicit research question. Recognized rules for planning and conducting clinical investigations for medical devices (including the preparation of study plans, etc.) can be found, for example, in good clinical practice, DIN EN ISO 14155:2021 [360], and the MDR.

Overall, the trial protocol should indicate whether an outcome is robust and meaningful for a specific endpoint (clinical or system) and select or develop a trial type (including trial protocol with transparent reporting) that provides the necessary evidence for an AI system by minimizing bias and provides confidence in the results. Ultimately, this can also provide certainty for decision-makers and users.

By all means, the basic principles of good clinical trial protocols can be applied in the same way to AI systems, taking into account specific requirements for their evaluation. So far, the level of trial design and reporting of published AI trials often does not meet the high standards in this regard (see e.g. [361]). For this reason, guidelines for improving specific trial designs in the evaluation of AI systems are being developed by various initiatives, such as at the international level by the interdisciplinary "EQUATOR Network" [EQUATOR Network.org], for the reporting of trial protocols e.g. "SPIRIT-AI" [362] or of trial reports "CONSORT-AI" [363]. In addition, an ethics vote must be included in the planning or implementation of a clinical trial. The results of the trial are to be reviewed by independent experts.

In the further course of the evaluation of the AI-based medical device, continuous testing of the performance and safety is necessary during the implementation phase of the AI model in order to register and remedy any unexpected effects that might only occur with use in a complex clinical environment. This also includes version updates of the AI. One way to address this is through AI audit procedures that can detect and accurately analyze unexpected effects, such as those described in [364].

User interaction
When using AI models in practice, interaction between users and the system is essential, as healthcare professionals are likely to adapt their own behaviour to automated support over time, e.g., by relying more and more on the system. This can lead to difficulties especially if the personnel cannot sufficiently understand the decisions of the system or if the system changes due to frequent new releases or even in the case of a continually learning system with constant adjustments. The users may then no longer be able to adjust sufficiently to the new system behaviour. In addition, the question must be answered as to which and how much information users need in order to understand and correctly classify the system's decisions. A response to this must additionally take into account the fact that users are often very heterogeneous in terms of their level of knowledge, their personal attitudes, or even in terms of the hospital environment associated with them.

Since such effects can often only be fully recorded in real operation, conclusive monitoring of the systems, also in conjunction with systematic recording of failure cases, but also positive results in the sense of post-market surveillance, is an important factor. The requirement for such steps is embodied in both the MDR and the proposed AI Act. Experience with the system is to be collected and evaluated in a targeted manner. In this regard, the proposed AI Act requires that human oversight maintains control of the system and is able to shut down (or switch to a classical mode) the system in a timely manner if necessary. As already mentioned, it is important that the users can achieve a sufficient understanding of the system and its decision-making basis, as well as a sufficient knowledge of the decisions based on it.

4.7.2.1 Use case: AI-assisted 2-D X-ray image analysis for caries diagnostics in dentistry

AI software applications are also increasingly being introduced into dentistry practices. The focus of current efforts is primarily on machine vision (computer vision), especially in the field of dental X-ray image analysis (diagnostic support), particularly in the 2-D X-ray range, e.g. analysis of single images, panoramic radiographs, bitewing images and cephalometric side images. This is due to the fact that a large number of radiographs are taken in dentistry (more than 50 million images in Germany and approx. 520 million worldwide per year), the accuracy of dentists in diagnosing on these images is limited (for example, the recall for detecting early caries on radiographs is < 50 %) and systematic and comprehensive reporting and documentation of the diagnostic results is costly.

Analyzing 2D dental radiographs using AI can help improve diagnostic accuracy, reliability, efficiency and communication of findings. AI-enabled medical devices must also undergo a thorough review process in this regard to ensure their safety, robustness, transparency, fairness, inclusivity, and (cost) efficiency.

The use case described below (see also Annex 13.5) of an AI component in dental 2D X-ray image analysis is in the context of digital intraoral X-ray imaging, specifically bitewing images. In this process, the posterior region of a patient's teeth is irradiated from the outside by means of a directed X-ray source and the signal is recorded by a digital X-ray sensor placed in the patient's oral cavity. This AI component is designed as a backend service: The created bitewing image as well as meta information (e.g. pixel size, radiation dose) are transmitted to this backend service, which runs on a cloud server. The execution of this backend service is thus initiated by the system itself and without human interaction. The data received is then automatically analyzed by a neural network trained in advance, and any detections of caries are output in the form of polygonal lines in the coordinate system of the digital bitewing image. Both (the unmodified digital radiograph and the points along the polygons of any caries detections) are provided via a secure network protocol to a workstation with software and appropriate user interface for reporting.

Gold standard
The most common method for AI-assisted 2-D radiographic image analysis in dentistry is supervised ML (see e.g. [365]); however, initial approaches have also used unsupervised learning [366]. In supervised learning, a gold standard (reference test) must be established by means of the annotation process.

There is no widely accepted gold standard for 2-D radiographic image analysis in dentistry; depending on the application focus (caries detection, detection of apical lesions, measurement of periodontal bone resorption), different reference test methods are used (including alternative imaging with high recall, e.g. 3-D radiographs such as digital volume tomograms or histological evaluations, e.g. on extracted teeth previously analyzed radiographically). For the field of AI applications, visual assessment of radiographs by dentists is common, and in the establishment of a gold standard, several experts are usually involved to address the imprecision of individual findings and to increase the validity of the gold standard. How exactly the gold standard is then constructed from the various findings is also not conclusively defined (see e.g. [367], [368]); for classification tasks, majority votes or consensus panels are used, e.g. [369]; for segmentation tasks, hierarchical procedures (three to five independent experts segment, a "master reviewer" reviews the totality of the segmentations) have been used, among others [366].

Describing and selecting the model

Convolutional neural networks (CNN) are primarily used for dental analysis of 2-D X-ray images. Depending on the task (classification, detection, segmentation), different model architectures as well as differently annotated data (see text passage "Gold standard") are used (cf. e.g. [365] or [370]). The model architectures used are based on the general state-of-the-art; no special architectures are used for dentistry.

For the training process, the data is split into independent training, validation and test datasets as usual; it is relevant to avoid contamination of the datasets (data snooping). This is particularly important in dentistry, as there are often several 2D X-ray images of one and the same patient. Splitting the datasets into training, test and validation datasets should therefore be done at patient level, not image level. Other parameters must also be taken into account when partitioning the data.

The selection of the appropriate model architecture is often empirical; there are hardly any systematic studies on the optimal model selection in dentistry (cf. [370]). Similarly, the determination of the hyperparameters during model validation (hyperparameter tuning) is currently performed mainly empirically (cf. [372]). The possibility of pre-training the models on radiographic datasets (e.g., freely available datasets containing lung radiographs) has already been demonstrated [366] and is an alternative to the usual pre-training on general, non-dental datasets.

CNNs are also used for the use case of caries detection in bitewing images. These are able to provide pixel-based classification, or "semantic segmentation". For this purpose, pixels located within marked polygons are assigned a numerical label to generate training, validation and test data; error metrics can thus be calculated on a pixel-by-pixel basis. In the practical use of such trained CNNs, the pixel-based outputs are converted to polygonal features by appropriate post-processing.

Performance criteria

To capture the performance of the models, metrics such as accuracy, recall, specificity, F-1, and the area under the receiver operating characteristic curve are determined for the test dataset. In addition to metric characterization, other quality criteria such as model robustness, fairness, "explainability," and the ability to adequately describe predictive imprecision should be considered.

Data management

The data collection for the training process must be designed so that the data are from the x-ray equipment to be encountered in the field according to the intended purpose, and they represent the population for the area of application. The annotations necessary for supervised learning must be made by qualified personnel, and their quality must be ensured by a review process (see text passage "Gold standard"). The training data is archived beyond the delivery of the product for documentation and for later repetitions of the training process. Typically, test data is fed to the AI model to assess performance as part of technical validation, and the model predictions are then compared to labels or annotations (defined by human experts as the "gold standard", see above).

AI-specific questions of a risk analysis

Risk analysis is guided by the general principles of risk identification (frequency, severity) and the derivation of mitigation strategies. In many cases, the most serious undesirable event is tooth loss; only for certain (e.g., surgical) AI applications on 2D radiographs is more extensive damage to be expected.

Clinical evaluation

In the clinical evaluation, the benefit of the AI system must be evaluated to see whether the AI works as intended in interaction with the users in the application environment, e.g. in the hospital or in a dental practice, and whether the medical staff, patients, etc. benefit from the method, i.e. whether it adds value. Questions relevant to this include:

→ Is the AI method safe and effective in the sometimes highly complex real-life application situations in dental practice?

→ Does the AI method provide measurable benefits beyond established methods?

Ideally, the clinical requirements should be considered from the very beginning, i.e. dentists should already be involved in the development process, and there should be a constant and ongoing exchange between developers, clinicians and all stakeholders involved. Clinical validation ideally involves conducting a randomized controlled trial or similar study design. In this context, aspects such as acceptance, implementability and maintenance, but also the influence on the diagnostic and therapeutic process (therapy decision) and the resulting cost-effectiveness of AI should also be considered.

User interaction
Within the AI software, treating dentists usually have the option of displaying the native image as well as the results of the AI analysis. Users of the system consequently use the results of the AI as an additional source of information during the diagnostic process (assistance). The detections displayed by the software can usually be deleted, edited, or new detections can be added by users and stored in the digital patient record alongside the image for documentation purposes. As described in the use case, the AI component is used for assisted caries diagnostics. The treating dentist can hide or modify the results of the AI component, delete false positive detections or manually add caries lesions missed (false negatives) by the AI component. Thus, all results are subject to human supervision

Summary
The performance, safety and efficiency of AI-based applications for analyzing 2D radiographs in dentistry must be ensured through standardization processes, among other things. In implementing standardization activities, the specific challenges in dentistry (including the presence of often multiple images of the same patient in a dataset; clustering of pathologies and statistical units at the patient and tooth level: patients and teeth sometimes have multiple pathologies whose occurrence is often not independent of each other; adverse health effects are often limited to the individual tooth) are taken into account, but will be largely analogous to other health fields.

4.7.2.2 Use case: AI-based artificial respiration system in intensive care medicine

While many current developments of AI-based medical applications (e.g. radiological examinations, image-based detection of eye or skin diseases) focus on the diagnostic aspect, intensive care applications increasingly include monitoring tasks or even aspects of controlling the course of therapy. As described in the context of the following application example, special requirements must be observed to be able to prove the safety and effectiveness of the system. This use case includes various gradations, such as the degree of autonomy (different levels of automation, the extent of human oversight or human-system interaction, see e.g. levels in autonomous driving, see also Chapter 4.7.3, Need for action 07-05 and [373]) and the timing of machine learning (one-time learning vs. continual learning incorporating new environmental data, see [374]). In the context of therapy support, it should be noted that not only pure classification or regression tasks are implemented by the ML component. Rather, it is a dynamically acting system that must repeatedly take measurements of the patients as part of monitoring or therapy management in order to be able to make predictions or adjust its own behaviour. The system interacts directly with the patients by supporting them in the course of therapy without the intervention of a treating person in the meantime. In this respect, it is a closed-loop approach (see [375]), although the degree of autonomy may also be limited, especially if the physician has to intervene at certain points. The description of this use case focuses on extended requirements that go beyond the use cases described so far.

Concrete use case: Ventilator with AI-assisted weaning
The primary purpose of ventilators is to supply patients with sufficient oxygen in the event of lung dysfunction and thus to assist them in breathing. This applies in particular to critical conditions, such as those that may occur after accidents or in the case of Covid with a severe course. The treatment involves a series of individual steps – from the decision to ventilation, to intubation and screening, to weaning and final extubation. In the context of the use case presented here, the focus is on ML-based weaning, in which ventilatory support is gradually reduced in order to eventually be able to switch it off completely when it is foreseeable that patient can manage their own breathing completely and permanently. The central therapy decisions regarding the initiation of the weaning phase or extubation must be made by a physician (trained specialist in anesthesiology/intensive care medicine) (see also Annex 13.5).

Weaning itself is to be controlled by an ML-based method by dynamically adjusting ventilation parameters based on continuous measurement of key physiological parameters and using trained neural networks to keep the patient in a stable state. In this process, weaning is achieved by gradually reducing respiratory support while assessing the current situation at any given time. It should be noted that during weaning, patients may enter different pathological states (such as hypo- or hyperventilation, tachypnea, ...), in which the system must respond appropriately. Accompanying alarms should be triggered to inform ICU staff of critical conditions and trigger necessary actions that are not subject to AI decision-making.

The system thus assumes both monitoring functions in conjunction with an alarm component for critical conditions and therapy support in the sense of a closed-loop system. Using a corresponding database from real cases, neuronal networks are to be trained in such a way that, on the one hand, they can recognize alarm situations and trigger alarms and, on the other hand, automatically implement necessary changes in ventilation parameters. Initially, a fixed data basis and a fixed state of the model for the conformity assessment are assumed. Extensions towards continual learning systems by adapting the neural network during operation to environmental parameters (such as the specific hospital environment) or to individual patient parameters are considered as an extension possibility.

Based on classical logic or physiological models, there are already systems on the market that realize such automated weaning and perform similar functions (identification of the patient's current state as well as the necessary support measures in the sense of a closed-loop system with an additional alarm component). However, decisions here are built on fixed criteria regarding key physiological parameters, such as spontaneous respiratory rate, tidal volume, and end-tidal CO_2. The systems must be operated by appropriately trained intensive care nurses, and again a physician must be consulted for certain steps (e.g., decision to start/stop weaning). These existing systems should be considered state-of-the-art/standard-of-care when developing an ML-based system.

A detailed description of this use case can be found in Table 19 of Annex 13.5. Below are listed special requirements related to the described use case, which complement the basic requirements listed in Chapter 4.7.2.

Special aspects of conformity assessment in the described use case "Intensive care"
On the technical side, a number of criteria must be implemented in the present case to ensure the basic safety of the system. Here we limit ourselves to aspects specifically related to the ML-based components of the ventilator, such as the measurement of the performance of the model, the provision/availability of appropriate data for training, and the testing (validation, testing) of the ML-based model in the present case. Ensuring the correctness and robustness (error tolerance) of metrics on which ML-based decisions are based, – (e.g. physiological metrics related to the patient's respiratory state) –, is also an important criterion in the present case in order to provide a reliable classification of system performance. Moreover, disclosure and evaluation of the basic clinical model on which the ML system is built (which physiological parameters require which clinical decision in terms of "appropriate" ventilatory support) is an essential factor for conformity assessment in such ML systems in terms of transparency. Another important issue as the level of autonomy of intensive care devices increases, is the ability of the operator to modify decisions of the ML system at any time to achieve better or more specific therapeutic outcomes or to take control in the event of an error.

When defining the evaluation criteria for the ML system, it is important to consider how they interact with actual clinical effects, i.e., what criteria characterize an optimal clinical outcome. Classical error criteria for training ML models such as accuracy, specificity or recall alone cannot directly classify how good the quality of the models is in a clinical context. For example, an overlooked alarm (false negative) can have significantly different effects than an incorrectly triggered and thus unnecessary alarm (false positive). The effects resulting from the individual error types should be specifically integrated into the error criteria in order to achieve a systematic minimization of the risk potential.

Moreover, in the case of ventilation control, the aforementioned classical performance criteria are not applicable, since it involves the optimization of a dynamic process in which performance must be measured in a different way. This effect is enhanced when ML models are combined with classical physiological models by, for example, hard-coding and integrating boundary conditions such as the response to known critical values in terms of spontaneous respiratory rate or tidal volume in a classical manner to avoid certain risks in the detection of pathological conditions. In such hybrid models, there is currently no clear guidance on how to compare existing classical and long-established approaches with responses from a novel ML-based system in terms of their clinical performance.

Furthermore, it should be considered how reliably the intensive care staff reacts to the alarms and how well they can classify alarms and understand their causes. Particularly in the field of intensive care medicine, reliable implementation and targeted testing of human-machine interaction with regard to its clinical effectiveness is an important step in order to reliably make decisions that potentially concern life-critical areas. The user must be specifically informed how reliable the results of the AI system are to be classified (e.g., via reliability scores for the alarms). It must be taken into account that users often rely either too much or too little on the system, which additionally alters its effectiveness. Ultimately, the system can only be fully tested in real-time operation, and there may be substantial differences in perception among different user groups and/or environments.

The ML-based determination of optimal decisions depends on many individual parameters (patients of different gender, age or ethnicity may have different breathing patterns) as well as on the complexity and possibilities of the respective environment (different equipment, different qualification of the care team, different approaches to treatment processes). The training, validation and test data must cover these scenarios as comprehensively and representatively as possible. The ML model itself must then capture as reliably as possible which action achieves the best clinical effect for which patients in which setting.

This requires better adaptation to the respective environment and patient population. On the other hand, a continual learning process would additionally be required to be able to adapt the ML system in an appropriate way. However, a high degree of flexibility in the models can in turn affect the safety of the system, especially in connection with user interaction. Users may not be able to adapt quickly enough to the new circumstances especially if the models are changed frequently. Therefore, control mechanisms should be integrated into the system to check such model drifts not only on a purely technical level, but also with regard to their clinical effect. On the one hand, in the area of human oversight, the proposed AI Act specifically calls for this. On the other hand, there are still no provisions for how this is to be implemented in the corresponding processes, especially for the safety-critical environment in the field of medicine or intensive care. In any case, the availability of suitable data from real-time operation is an important factor in achieving high performance. To this end, extensive usability of real-time data and incremental integration of data into the model, as well as the testing required to achieve this, is a key aspect for the development of AI-based applications in the field of critical care medicine.

Summary

Overall and in light of its risk aspects, the present use case is a rather complex and critical scenario, which is not yet sufficiently covered in existing standards (e.g. DIN EN ISO 14971:2022 [351] with regard to risk management, DIN EN 62304:2016 [353] with regard to the software life cycle, DIN EN 62366-1:2021 [355] with regard to usability/ human-machine interaction) (see also [356]). A number of aspects are highlighted in the scenario that play a subordinate role in purely diagnostic use cases, but which should be addressed in the future in order to obtain comprehensive provisions for the implementation of AI-based systems in the various application scenarios.

4.7.2.3 Use case: Segmentation and classification of brain areas (including cerebrospinal fluid) and their volume determination

State of the art

Magnetic resonance imaging (MRI) has established itself as a standard procedure in neuroradiological diagnostics. In particular, it can be used to visualize different tissue structures in relation to each other and pathological changes in normal tissue. In addition to the visual evaluation of 3D MRI data, quantitative measurement of anatomical structures and, if necessary, their changes over time, e.g., to monitor the course of therapy, is often necessary for diagnosis. For a conventional volume determination, selected anatomical structures must be segmented / marked in the image manually or semi-automatically (e.g. by contrast or edge detection) for this purpose. However, due to the time-consuming nature of this process, it is rarely performed in everyday clinical practice. As a substitute, simple length measurements are often taken, the diagnostic significance of which is generally significantly limited compared to volume determination.

Concrete use case: Segmentation and classification of brain areas (including cerebrospinal fluid) and their volume determination

In the use case described, an AI-supported, fully automated segmentation of all relevant brain areas takes place. For this purpose, the 3D MRI images of a patient's head are sent to a central picture archiving and communication system (PACS) and analyzed there by a radiologist. The 3D MRI images are automatically analyzed, i.e. specific regions are volumetrically quantified and then visualized for radiologists (e.g. in a clear report containing precise quantitative information and highlighting specific lesions). The current use case is limited to an AI component that supports the diagnosis of neurological diseases in radiology. This AI component runs on a separate data processing computing unit that is locally integrated into the radiology infrastructure. The calculation is executed by receiving the data from the PACS. The output contents (reports and visualizations) are restored to the PACS after completion of the calculation and are made available to the radiologists together with the acquired 3D MRI images (see also Annex 13.5).

The volumetry report can be used to support diagnoses in neurodegenerative diseases (such as Alzheimer's disease, frontotemporal dementia, multiple sclerosis, and forms of Parkinson's disease). Radiologists can display the 3D MRI images either alone or together with the results of the AI component. Consequently, users of the system use the results of the AI component as an additional source of information during diagnosis and can thus supplement the purely qualitative diagnosis with quantitative information and additional visualizations. The established diagnostic findings are not replaced by results from the AI component, but enriched by the supplementary information generated. By visualizing the AI-based results, radiologists are able to evaluate the correctness of the information generated by the AI component. The matching of the structures segmented by the AI with the anatomical reality present in the image data is performed within the professional competence of the clinical users and does not require any specific training with regard to the AI component.

Describing and selecting the model

Segmentation and classification of regions in the brain first requires processing of the data to create the 3-D datasets of the brain. For semantic segmentation of brain regions, CNNs are used in the form of encoder-decoder architectures. In this regard, the state of the art in science offers a variety of neural network architectures for semantic segmentation based on 3-D datasets such as the DeepLabV3+ [376] or the U-Net [377].

The training and validation process must be explained in addition to describing the architecture of the AI model used. Which selection of model parameters (hyperparameters) and preprocessed datasets led to the desired prediction result? If subsequent post-processing becomes necessary, this must also be described.

Performance criteria

A common performance criterion for evaluating the quality of a three-dimensional semantic segmentation in the AI environment is the intersection over union (IoU). This metric represents the numerical ratio of the intersection to the union between predicted and actual segmentation. The result is evaluated voxel by voxel. Another criterion is the F1 score or dice coefficient and mean average precision.

If the result is needed by the user at short notice, such as during an operation, the time until the AI result is available must be taken into account and must also be validated.

Data management

As mentioned in the general Chapter 4.7.2, there is a need for a common understanding of which demographic, epidemiological, and indication-specific variables have a significant impact on medical purpose and, consequently, need to be represented appropriately in the data. Certain demographic attributes may be included in DICOM data generated and used by imaging systems. However, it should be noted that this data is not always available or that there is sometimes no DICOM attribute to capture corresponding data. There is also a need for clarification as to what level of detail demographic attributes may be used for what purpose (e.g., training machine learning models) in compliance with the GDPR.

When labelling or annotating training data for supervised learning, general aspects need to be considered, such as the qualification of personnel and the use of validated software tools. This applies both to the generation of the annotations and to the checking of the annotations by a second person. This use case results in specific requirements for the software tools used for labelling: The software should be able to process metadata and ideally have features that support the annotation process and make it more efficient, such as automated determination of the field of view relevant for annotation to reduce the time spent scrolling and zooming. In addition, an evaluation or annotation of the quality of the image dataset is recommended as part of the annotation and review process.

Another aspect is the use of synthetic data to enrich the dataset with new features. Generative adversial networks (GAN), which have already been applied to tumour segmentation in the brain based on MRI data, represent an interesting approach in this context [378].

Risk management
For the described use case, the risk management procedure is to be classified in particular with regard to the function in the context of the mediated image-based measurement function. The following specific challenges arise.

Identification and evaluation of errors that occur in the calculation of the volumetrics report due to the use of an MRI. Variations are possible between the individual slice images of the MRI in each case. Also, the quality of the calculation of the volumetrics report must be considered with respect to effects due to age, gender, ethnicity as well as pre-existing conditions. Varying density or change in the tissue being measured due to scarring or swelling from brain inflammation, brain aneurysm, infarction, or obesity may have an effect on the calculation of the volumetrics report. In addition, the presentation should allow a checking of the usability of the results. The fact that users can directly assess the anatomical correctness of the segmentations generated by the AI with appropriate visualization based on their clinical knowledge ensures the application safety of the solution.

Clinical evaluation
The software is intended to provide information used to make decisions for diagnostic or therapeutic purposes (volumes, anatomy) and does not perform diagnosis. The clinical evaluation should be performed according to a defined and methodical procedure. If sufficient robust scientific literature is available to allow evaluation of the safety, performance, and design characteristics of the product, a systematic review can be used to compile the evidence for the above points.

In the case where the AI component is a new technology for which there is no or not enough scientific literature, a clinical trial is necessary to collect enough data on the safety and performance of the product, especially to prove the performance of the AI component. Such trials should produce enough data to allow a generalizable statement about the segmentation of brain regions and their volume values.

The primary objective of such trials would be to prove the performance of the product within its intended purpose. To achieve this, a comparison between algorithmically and expertly determined volume values should take place. One possible measure of interrater reliability is Cohens-Kappa. For this, the following literature source requires a minimum value of $\kappa = 0{,}4$, where $\kappa > 0{,}6$ counts as substantial and $\kappa > 0{,}8$ counts as an excellent matching result [379], [380].

The clinical evaluation report should support the safety and performance of the product. For this purpose, the results of the trials as well as the non-clinical data generated by the non-clinical test methods (e.g. by usability, verification) are used.

User interaction
As described, the AI medical device is intended to support clinical diagnostics. The responsibility of diagnosis rests with clinical staff trained on the AI component. The results of the AI software can be added directly to the patient's digital record, saving time in the clinical day-to-day work.

In order for the AI medical device to add value to everyday clinical practice, it is necessary to process the result. Since brain volume (or the volume of individual regions) changes with age, a comparison with the age cohort would be useful, as well as a comparison of the volume of a patient at different time points.

Summary
The described AI-supported use case for volumetry of brain regions and cerebrospinal fluid is intended to support physicians in making diagnostic decisions about neurodegenerative diseases. The results determined by the AI are summarized in a report and can be visualized. A useful addition would be the direct comparison of the determined volume values with, for example, the average age cohort or a progression diagram for tracking changes in brain volumes in individual patients.

To be able to improve the software in a timely manner, the possibility of a feedback loop for false detections by clinical staff would be desirable. Such a need can be met by a continual or incremental learning system, which is discussed in more detail in Chapter 2 under Recommendation for action 4.

4.7.3 Standardization needs

Need 07-01: Usability of data for AI-based systems in medicine
AI systems have a high need for data in order to be able to derive reliable statements and also to enable reliable evaluations regarding their performance in terms of validation. In the medical field, data has some special requirements that need to be considered as part of the development and validation of AI-based systems. First, it typically involves **personal data** that must meet strict data protection requirements.

Second, there are high standards for data collection in the medical field (implementation in dedicated study designs, involvement of an ethics committee, high standards with respect to statistical analysis), which means that access to data is additionally limited. Because the MDR often requires clinical trials to be conducted in order to bring a product to market (i.e., elaborate and costly studies in an often relatively restricted context based on a not-yet-approved product), it specifically limits the collection of data that come from real-world and wide-ranging use environments. Therefore, the third question would be what other sources of data would be permissible (e.g., real world data from similar applications or operation of an existing system, benchmark data provided, accessible central databases such as envisioned in the proposed EHDS, synthetic data) and how they could be acquired or used to have more data available from real-world environments. And fourth, it would have to be clarified to what extent the existing statistical requirements established with respect to classical study designs have to be adapted if instead of individual dedicated parameters, machine learning methods often have to cover much more complex scenarios. This would also raise the question of whether a randomized controlled trial should still be considered the benchmark for AI-based systems or whether alternatives would be more appropriate here (e.g., for updates to an existing system using real world data). The overriding requirement here is to clarify when a dataset and selection can be considered sufficiently representative for a particular application and which data can or even must be used in which way for this purpose.

In summary, the following points need to be clarified and implemented in standardization as needs for action:

→ Clarification of requirements for data management and associated processes, including specifications for data acquisition, labelling, qualifications of individuals involved, and an overall standardized assessment of data collections.

→ Clarification of requirements for (clinical) study designs that are usable for validation of AI-based medical devices – including the framework of the design, the scope of the data, representativeness for the use case, access to the data (e.g., by notified bodies for review as required by the planned AI Act), usability of data from central databases, and framework conditions with respect to data protection.

→ Clarification of requirements for the extent to which real-world data can be used to develop and test AI-based medical devices. This potentially includes the use of co-logged data from the operation of an existing version of the system, but also the use of another system that systematically collects data from the operating environment.

→ Clarification of requirements for the use of synthetic data for AI-based medical devices. This includes the application of generic procedures in the context of machine learning, such as GANs, as well as specific models to generate new data, such as an artificially generated transformation of an MRI dataset to a different parameter representation or a new procedure with reference to a deterministic model for the transfer from the existing to the new modality. In the medical context, the use of synthesized data requires clarification of the extent to which compliance with specific requirements (e.g., the reliability of the generated data for the respective application, data protection aspects) must be demonstrated.

Need 07-02: Design of suitable metrics for different types of AI-based medical devices

Conformity assessment for medical devices requires a systematic review of their performance and also their safety, which must be appropriately quantified for AI systems in the form of suitable metrics. Here, there are some differences compared to other sectors. In the case of medical devices, the clinical outcome, which includes both risks and benefits for the patient, must always be evaluated. Therefore, in clinical evaluation, the MDR requires a systematic consideration of the risk-benefit balance as a central step in the conformity assessment process. In addition, a comparison with reference procedures such as the established standard of care and also the products already on the market is necessary. This requires that defined reference criteria are available, not only to evaluate individual systems, but also to implement a targeted comparison between systems, including specifications as to when systems can be considered equivalant. Ideally, benchmarking datasets should be available for different applications in order to be able to perform a standardized comparison.

In addition, in the medical field, assessments must be implemented in a highly use-case-specific manner so that the specific benefits or resulting risks can be assessed in a targeted manner. It must be taken into account that there are very different areas of application, which include, for example, tasks in the areas of diagnostic, monitoring and therapeutivc procedures, and that these can additionally be associated with different degrees of autonomy and risk. In the case of AI-based systems in particular, factors such as transparency or explainability (how does the AI come to which decision, which basic assumptions does it use as a basis, which steps does the AI perform at the current time) as well as the possibilities for intervention within the framework of human oversight can be included as criteria in addition to technical risks. These should also be considered in terms of their clinical efficacy.

To effectively reduce risks and optimize benefits, it is basically necessary to incorporate these factors into the evaluation metrics. Such factors can often be conflicting (e.g., transparency vs. accuracy, safety for individual patients vs. benefit for a particular population), creating trade-offs that the evaluation criterion as a whole must capture. Reducing individual risks, as typically targeted in risk management standards, is only partially effective, especially for AI-based systems. It should be noted that the integration of risk aspects into the metrics should be done in a tiered manner, as often it is not possible to fully quantify these points during development. This integration should not be an essential hurdle to the successful assessment of a medical device's conformity, but should maximize the systems' potential for improvement as specifically as possible.

In this respect, there are some requirements that go beyond existing approaches and for which AI-specific specifications need to be developed. It needs to be clarified which metrics are relevant with respect to AI-based systems and how these are to be implemented with respect to conformity assessment. This includes the following aspects in particular.

→ The provision of standardized metrics to implement a systematic comparison of different systems for comparable use cases.
→ Integration of AI-specific risk factors and aspects of clinical benefit into the evaluation criteria so that optimization of the risk-benefit ratio in the overall system can be implemented in an appropriate manner.
→ Establishment of verifiable, possibly tiered requirements for transparency and explainability that allow users and patients to understand the basic operating principle, while at the same time providing guidance to the user in critically evaluating AI-based decisions.
→ Consideration of different degrees of autonomy and areas of application (e.g., diagnostic vs. monitoring vs. therapeutic procedures) and the quality criteria associated with each. Clarification of their interaction with measures addressed in the development and life cycle process of medical devices.

Need 07-03: Societal and regulatory framework for the use of AI in medical devices
The development of the MDR or the In Vitro Diagnostic Medical Devices Regulation (IVDR) has shown how important it is to design regulations in such a way that they can be easily implemented within the given time frame and have a positive impact on healthcare. Excessively high hurdles (e.g., need for clinical trials for long-established existing products with limited risk potential) and unavailable infrastructure (e.g., availability of central databases such as Eudamed, lack of notified bodies and harmonized standards) not only lead to uncertainties in the development process, but can also cause problems for Germany as an industrial location and for healthcare as a whole (e.g., lack of important niche products).

The planned AI Act of the EU adds an additional level of complexity, which involves interactions with existing regulations and can thus result in additional burdens with regard to the conformity assessment of AI-based medical devices (see also Annex 13.1, section "Exemplary presentation using the example of medical devices"). To avoid disproportionately limiting innovation in this important area of the future, inconsistencies between the proposed AI Act and the MDR (e.g., in the area of risk management, post-market surveillance databases) should be eliminated and duplication of effort minimized. Accompanying the design of the planned AI Act, the relevant standards should be prepared to ensure that the requirements are implemented consistently and as efficiently as possible.

In addition, accessibility to data should be improved to avoid setting disproportionately high barriers to the development and implementation of new AI-based applications in medicine. A good balance is needed here between the data protection aspects given by the GDPR and the medical technology requirements that need comprehensive data to achieve a suitable level of safety or clinical benefit. Efforts toward the EHDS are one approach in this direction. It should be noted, however, that companies also need access to appropriate data in order to be competitive and innovative.

Overall, the focus in the future should be more on evaluating the extent to which new products have a positive effect on the entire healthcare system. This means that it should be possible to weigh the risks of an individual product more closely against the benefits for society as a whole. A product that is not available on the market due to excessive hurdles and is therefore absent from the healthcare system also causes harm. In this context, there should be a targeted evaluation that assesses the impact of the regulations themselves (and the associated standards) not only in terms of their technical implementation, but also in terms of their impact on the entire healthcare system.

Specifically, the following needs arise:

→ **Eliminate inconsistencies and duplications between the planned AI Act and the MDR** or IVDR (e.g., in the area of risk management, post market surveillance databases).

→ **Ensure that the infrastructure for the implementation of the planned AI Act is prepared**, e.g. with regard to the detailed clarification of the requirements contained therein, provision of harmonized standards, availability of central databases (e.g. for post-market surveillance), and availability of notified bodies certified for both the planned AI Act and the MDR or IVDR. Care must be taken to ensure that the transition periods of the proposed AI Act are designed accordingly.

→ Improve access to medical data in order to strengthen innovation in Germany/Europe and also improve healthcare overall via AI-based systems. This should include appropriate access to data for companies.

→ Greater inclusion of the positive effects of (AI-based) medical devices in the context of conformity assessment or risk-benefit assessment.

→ Targeted evaluation of regulations and standards (on a scientific and independent basis) in terms of their impact with regard to healthcare as a whole.

Need 07-04: Degrees of autonomy in AI-based systems – different levels from human-in-the-loop to closed-loop models

In the field of AI-based applications in medicine, there is a wide range of degrees of autonomy that occur in different tasks – from a pure logging of data to dedicated diagnostic decision support and support systems in the field of monitoring (such as intensive care) to highly automated systems. With a low degree of autonomy, users (e.g., medically trained personnel) must be able to monitor algorithmic results in a reliable manner ("human/clinician in the loop" systems). This requires that users have a sufficient understanding of the system, even in dynamic and complex environments, to be able to react correctly to its decisions. In a highly automated approach – in extreme cases a closed-loop system – on the other hand, the central system behaviour must be controlled without human intervention and still function safely. In contrast to other sectors (e.g. automotive sector with gradations from assisted driving to autonomous driving), there is no consistent classification into degrees of autonomy in the field of medical technology, but only very limited points of reference. For example, PD IEC/TR 60601-4-1:2017 [373] includes classification tables for autonomy levels in Annex C, but these do not address AI-based aspects. For closed-loop systems that use classical, rule-based approaches (e.g., based on physiological models), there is a normative basis in DIN EN 60601-1-10:2021 [375], which, however, also focuses on the classical physiological control loop rather than an AI-based system.

Therefore, there is a need to clarify which degrees of autonomy are relevant with respect to AI-based systems and how they affect conformity assessment for AI-based medical devices and specifically for interaction with humans (for human-in-the-loop systems) or with physiological systems (for some closed-loop systems). This includes the following aspects in particular.

→ Definition of different levels of autonomy and clarification of the resulting requirements with regard to measures in the development and life cycle process of medical devices. This concerns in particular the influence of the levels of autonomy with respect to the assessment/treatment of risks, the validation of the systems or even monitoring in the field. In addition, the interaction with other risk assessment parameters, e.g., severity and probabilities of occurrence, must be considered. For ML-based systems, other parameters such as complexity and interpretability of the systems are added. Overall, a consistent risk-based approach must be developed that takes appropriate account of the gradations with regard to the levels of autonomy and enables adjustments to the associated requirements in a targeted manner.

→ Especially in human-in-the-loop approaches: Clarification of requirements related to human-machine interaction (see also requirements for human oversight in the proposed AI Act): What information do users need and in what way in order to be able to implement necessary reactions, e.g. in differentiating between alarms (immediate necessity of an action) and alerts (alerting to initiate further clarification steps)? This also includes clarifications regarding the requirements for transparency and explainability/interpretability of the systems, especially with regard to the very dynamic system behaviour that AI-based systems can exhibit. This involves clarifying what measures human oversight can include to avoid risks such as over- or under-reliance on the system's decisions or model drift.

→ Especially for closed-loop systems: Clarification of reliability requirements for AI-based systems or components that are not based primarily on established physiological models, as previous closed-loop systems have been. In addition, clarification of when a system represents only a configuration of parameters (e.g., AI-based estimation/adjustment of individual parameters) and when it represents a change in closed-loop system behaviour, as well as clarification of under what conditions/with what requirements combinations of AI-based and classical physiological models, i.e., hybrid models, can be placed on the market.

Need 07-05: Clarification of the distinction between medical and non-medical devices in conjunction with tiered requirements for AI-based systems in the healthcare sector
AI-based applications in medicine can cover a wide range of systems and components. First, there is the area of medical devices, the requirements for which in Europe are governed by the MDR and for which, in many cases, extensive additional requirements will be added from the proposed AI Act, now that the MDR is considered a substantive indicator for classifying an AI-based system as a high-risk system under the proposed AI Act. Second, there are the in vitro diagnostic devices, whose conformity assessment in Europe is regulated by the IVDR. Due to the relationship between the MDR and IVDR and the fact that IVDR products are also explicitly listed as candidates for high-risk products, requirements similar to those for medical devices apply to IVDRs to a not insignificant extent.

In addition, there are AI-based health-related applications that are not covered by the MDR or IVDR. This includes general health applications to individually manage, maintain or improve one's health (e.g., through "smart watches" or other applications that measure health parameters and process them in an AI-based manner). Specific regulations or standards currently only have a very limited impact here. As an exception, DIN EN 82304-1:2018 [354] provides some general requirements, but it does not include specific aspects of AI systems. Since such tools will gain further importance in the future and they also have to work reliably in order to avoid risks and to be able to have a positive effect on the health of the individual, specifications for the implementation of such systems would be helpful. This would need to be addressed in particular by providing suitable standards. In order not to unduly restrict innovation, these requirements should be differentiated from the stricter requirements for medical devices. In this context, it would be advantageous if agreement could be reached at the international level as to when an AI-based system is a general health application and when it is a medical device, or what classification in terms of risk level the respective system has. The same applies to healthcare systems that do not themselves have a medical purpose and therefore do not constitute a medical device, but which support processes in the healthcare system, e.g. optimization of processes in a hospital or care facility.

A final type of AI applications are components and tools that support the development, conformity assessment, or operation of medical devices, e.g., in the area of quality assurance, optimization of processes and products, or evaluation in the area of post-market surveillance. To a certain extent, these components fall within the sphere of influence of the MDR or IVDR and must, for example, fulfil corresponding requirements in accordance with DIN EN ISO 13485:2021 [381] ("computer system validation"). However, again no AI-specific aspects have been covered to date, so the current standards should be supplemented accordingly.

In all cases of non-medical devices, the problem remains in the proposed AI Act that while it regulates high-risk AI systems in great detail, it includes little clarification of how systems should be placed on the market that involve lower risk requirements.

In summary, the following key needs for action remain:

→ **Improved demarcation between medical devices and non-medical devices or consistent classification of the respective AI-based systems with regard to their risk level** – ideally in an international consensus, which shall, however, be coordinated with the respective legislative requirements.

→ **Definition of reduced requirements** for AI-based systems or even subcomponents that should themselves be lower risk, but still include high reliability to have a positive effect on healthcare. These include, in particular, general healthcare applications, AI-based systems for improving healthcare facility processes, and tools for developing and optimizing medical devices and in-vitro diagnostics.

Need 07-06: Application of assurance cases to provide safety evidence for AI-based applications in the medical field
As an alternative to the interpretation of existing rule-based standards that insufficiently address the field of AI, a more goal-oriented approach to safety verification of AI components using the concept of assurance cases defined in ISO/IEC/IEE 15026-1:2019 [114] appears to be a sensible basis and bridge to upcoming AI standards [382], especially in the medical field, and should therefore be considered more intensively. In this standard, an assurance case is understood here as a justifiable and verifiable artefact that supports the

assumption that an assertion made (e.g., regarding the safety of a medical device) is satisfied, comprising a systematic argument and the underlying evidence and explicit assumptions on which the assertion is based as summed up in ISO/IEC/IEE 15026-1:2019 [114].

The use of assurance cases is particularly recommended when innovative use cases are to be implemented or new technologies are to be used [383]. Both are usually the case when AI is used in medical devices. By means of assurance cases, the evidence of compliance with risk acceptance criteria accepted in the respective field (cf. e.g. [384]) can thus be broken down in a structured manner to the evidence provided by quality assurance [385]. This makes it possible to transparently justify the relevance and contribution of the respective measures in securing the AI portions of the product from ensuring an acceptable risk-benefit ratio as well as reducing the residual risk.

Experience from the use of assurance cases as structured argumentations also supports the development of standards with justifiable requirements. The development and consolidation of suitable argumentation templates in the use of AI in medical devices as well as their practical application, for example in the context of experimental rooms, should therefore be promoted by politics.

Recommendations:

→ Promote the use of assurance cases as a meaningful foundation and bridge to upcoming AI standards.

The Working Group Medicine ranked the identified needs according to the urgency of their implementation. Figure 43 shows the urgency of implementation, categorized according to the target groups of standardization and policy.

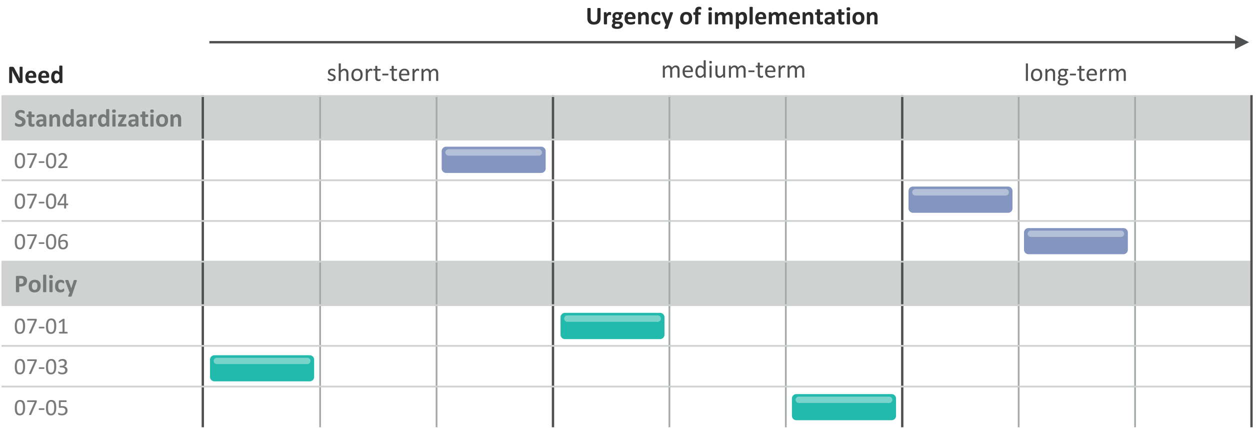
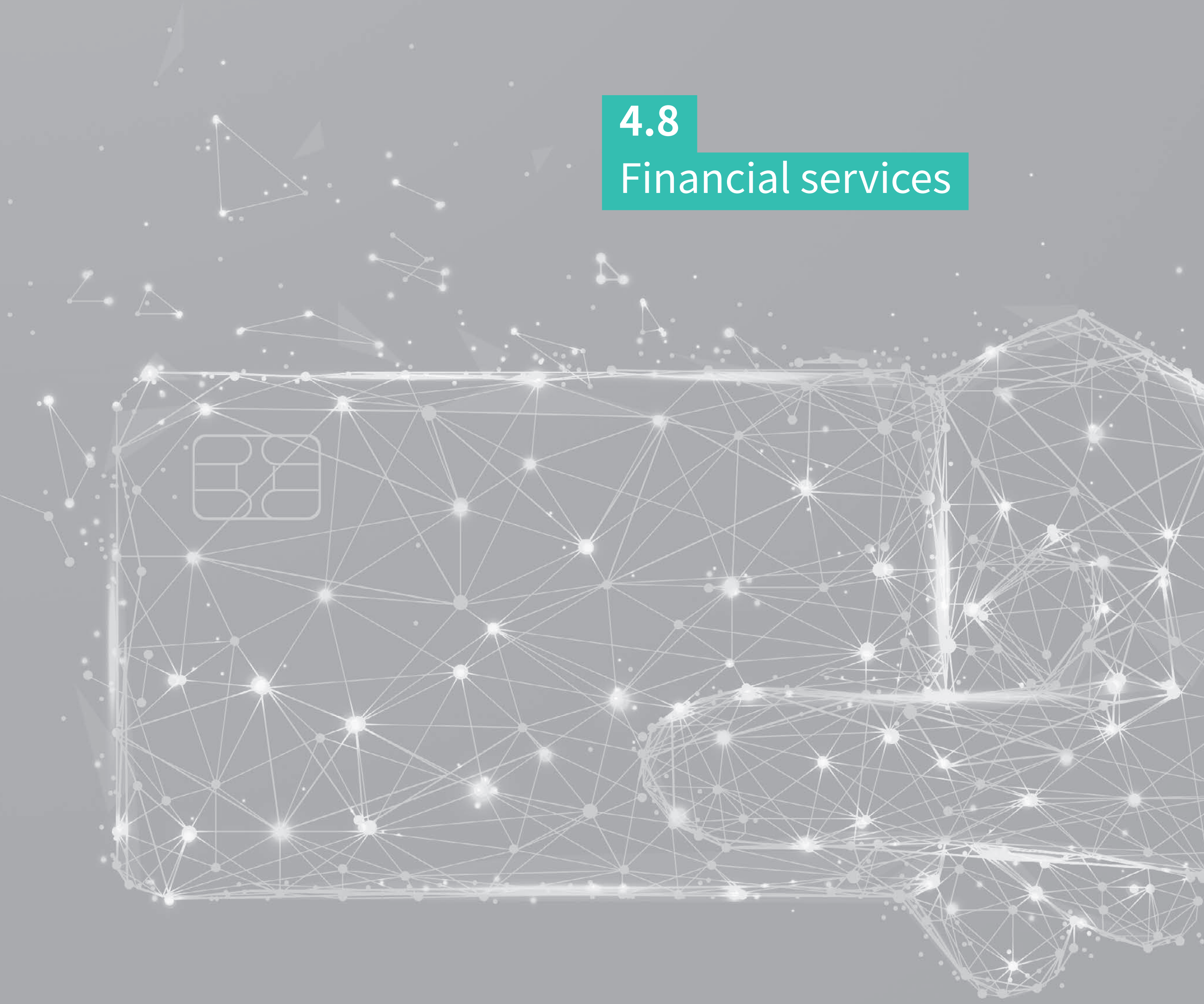


Figure 43: Prioritization of needs for the key topic medicine (Source: Working Group Medicine)

4.8 Financial services



Artificial intelligence (AI) is one of the key technologies of the 21st century that will influence business activities and the lives of citizens in many ways in the coming years and decades. With regard to businesses, especially financial institutions, not only are business processes being automated with AI, but entirely new business and operating models are also being developed. This can lead to businesses organizing and creating value in a completely new way.

However, AI also harbours risks that come to light particularly with regard to society. Leaving critical decisions to an AI system can result in certain populations being treated unfairly or even discriminated against. This must absolutely be avoided. The protection of fundamental rights must be the top priority when using AI systems. Citizens of the European Union (EU) or people living here must be protected from the risks of AI systems in all areas of life.

The finance industry already relies on AI and is unimaginable without the use of AI systems, see Figure 44. Financial institutions are facing a continuous growth of data with which they must deal. In addition, AI holds opportunities for innovation that cannot be foreseen now. Start-ups or FinTechs in the finance sector show that completely new business models can be developed with AI, but the business models of established banks are also undergoing a transformation shaped by data-driven systems.

4.8.1 Status quo

AI can be used in financial institutions to automate a range of use cases (see Figure 44):

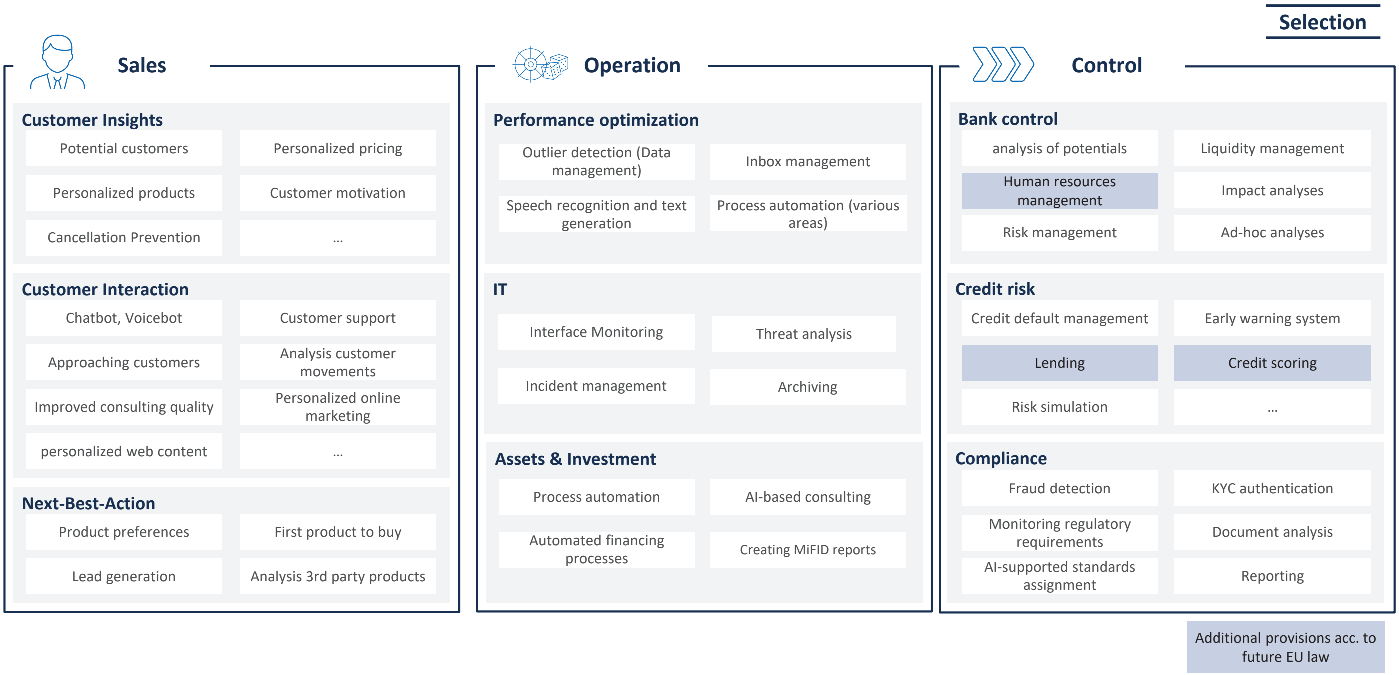
Fundamentally, the use of AI systems poses three risks that must be addressed by financial institutions:

→ **High complexity:** The machine learning (ML)-generated models used in modern AI systems can have significantly higher complexity than was the case with classical AI systems, making it much more difficult to track and test the systems.

→ **Short recalibration cycles:** Since the ML models used in AI systems are retrained at shorter intervals and thus constantly evolve, validation must also be dynamic.

→ **Bias:** The risk of biased results and unfair treatment of certain populations increases due to potential biases in the large and complex data sources used that are difficult to detect.

The EU Commission's draft AI regulation identifies several use cases that pose high risks through the use of AI, two of which are relevant to the financial industry: Credit scoring for lending and employee management systems. The EU Commission's argument here is that bias (i.e., a systematic bias in output against a known correct outcome, which can have



Selection

Sales

Customer Insights
Potential customers | Personalized pricing
Personalized products | Customer motivation
Cancellation Prevention | ...

Customer Interaction
Chatbot, Voicebot | Customer support
Approaching customers | Analysis customer movements
Improved consulting quality | Personalized online marketing
personalized web content | ...

Next-Best-Action
Product preferences | First product to buy
Lead generation | Analysis 3rd party products

Operation

Performance optimization
Outlier detection (Data management) | Inbox management
Speech recognition and text generation | Process automation (various areas)

IT
Interface Monitoring | Threat analysis
Incident management | Archiving

Assets & Investment
Process automation | AI-based consulting
Automated financing processes | Creating MiFID reports

Control

Bank control
analysis of potentials | Liquidity management
▨ Human resources management | Impact analyses
Risk management | Ad-hoc analyses

Credit risk
Credit default management | Early warning system
▨ Lending | ▨ Credit scoring
Risk simulation | ...

Compliance
Fraud detection | KYC authentication
Monitoring regulatory requirements | Document analysis
AI-supported standards assignment | Reporting

▨ Additional provisions acc. to future EU law

Figure 44: AI in the finance sector (Source: Working Group Financial Services)

different causes) in the AI system makes it more difficult for certain groups of the population to access financial resources or further development and advancement opportunities in their careers. Therefore, certain control mechanisms must be put in place here to prevent unfair treatment of certain people or to correct it after the fact. Since data often simply reflect the realities of society, existing negative trends and the disadvantaging of certain population groups are also adopted. In addition, biases can arise from learning data being biased through gaps in key aspects.

Nevertheless, financial institutions are equipped with very complex risk management systems, enabling companies to identify, prevent or mitigate new risks. In addition, the risks mentioned here are already known to the financial institutions and are addressed and controlled by the risk management processes.

4.8.2 Requirements and challenges

This chapter will address the specific requirements that arise from the application of AI in the finance sector. The specificity arises mainly from the following two circumstances, on the one hand from the consumer's perspective, on the other hand from the institution's perspective.

From the consumer's perspective
From the customers' perspective, AI applications in the finance sector are often applications with a direct link to people, similar to applications in the field of sociotechnical systems. The following aspects pose specific challenges:

→ Models for human behaviour are often needed here. This behaviour is usually variable, time-varying and strongly interconnected between the individuals whose behaviour the model describes.
→ Complex, interconnected, and unstructured data (i.e., data that requires interpretation) must often be considered to build these models.
→ Models are often created for groups whose average properties are to be transferred to individuals.
→ The results provided by the models may be related to important preconditions of social participation, including fundamental rights, such as access to credit and other basic financial services.
→ Due to the strict confidentiality and need for protection of personal financial data, there is a particular relevance of data protection and data security.

From the institution's perspective
Financial resources, especially in the literal sense of the word loans, are inherently risky, the business environment of financial service providers is strongly characterized by risk and complexity, and interdependencies between systems have to be taken into account. Risk management is therefore part of the core business model of any financial firm, and both the use of ML-based AI systems and the management of the associated model risks have long been established standards in financial institutions, which are also constantly monitored by regulators. This results in special challenges when dealing with AI systems:

→ The already existing system landscape of AI models is an integral part of the business model; interventions, also due to new standards, can therefore have a major impact on the capital base, for example. Therefore, all terminology must be precisely defined and standards quantitatively well justified.
→ New validation and certification processes must be embedded in the existing model risk management framework.
→ The fundamental focus from an institution‘s perspective is on managing portfolio risks, not individual risks, in line with regulatory requirements.

With this in mind, the five key topics for AI standardization in finance described below were identified.

4.8.2.1 Special features of the financial sector

Legal and regulatory requirements for the financial industry
Few industries are subject to as many regulatory requirements as the financial sector, especially the banking industry and payment transactions, compliance with which is strictly monitored by sector-specific supervisory authorities at national and European level.

The regulatory requirements that are particularly relevant to AI systems are those related to the risk management system and information technology of financial institutions. It is important to mention here that banks' risk management systems are capable of identifying, assessing and mitigating new risks so that risks arising from the use of AI systems can be managed appropriately.

The following is a summary of the requirements that apply to information technology systems in the financial sector.

European requirements

DIGITAL OPERATIONAL RESILIENCE ACT
On September 24, 2020, the EU Commission published the draft for the "Regulation on the digital operational resilience for the financial sector [386]". The regulation aims to ensure that all participants in the financial system have the necessary safeguards in place to prevent and mitigate cyberattacks as well as other risks. Financial regulators are to have access to information about IT-related incidents and ensure that financial firms are paying attention to the effectiveness of their preventive and resilience measures and are identifying and resolving vulnerabilities.

National requirements

SECOND ACT TO INCREASE THE SECURITY OF INFORMATION TECHNOLOGY SYSTEMS (GERMAN IT SECURITY ACT)
As a "critical infrastructure", the financial sector is subject to the scope of the "Second Act to Increase the Security of Information Technology Systems" (IT Security Act), which came into force at the end of May 2022. The act aims to ensure that the use of certain IT components by operators of critical infrastructures can now be prohibited if it can be assumed that the use of these components is likely to impair Germany's public order or security. The exact requirements are further specified by the "Ordinance on the Designation of Critical Infrastructures". Every deployment of critical components of critical infrastructures must be reported to and reviewed by the German Federal Ministry of the Interior.

MINIMUM REQUIREMENTS FOR RISK MANAGEMENT FOR GERMAN BANKS (MARISK) AND BANK SUPERVISORY REQUIREMENTS FOR IT (BAIT)
The MaRisk contains detailed regulatory requirements from the German Federal Financial Supervisory Authority (BaFin) for the design of the risk management system of German banks. The MaRisk is a set of administrative regulations that interprets Section 25a (1) of the German Banking Act (Kreditwesengesetz – KWG). The aim is to design the risk management system appropriately and document it in a meaningful way. All critical areas of a lending institution should be monitored by the risk management system, including the lending process.

Since IT provides the basic infrastructure for all processes of a financial institution, the regulatory requirements for IT have been elaborated in the Bank Supervisory Requirements for IT (BAIT). The aim of the BAIT is to ensure that banks' IT systems are appropriately equipped from a technical and organizational point of view, and focuses, among other things, on the requirements of information security and emergency management.

A risk management-specific consideration of the legal requirements can be found in Chapter 4.8.2.5.

Furthermore, the following standards and regulations are relevant, among others:

- → IT requirements for additional financial service providers: see Chapter 4.8.2.4.
- → IT security: BSI-Kritisverordnung (Critical Infrastructure Ordinance), IT-Grundschutz of the BSI (German Federal Office for Information Security), ISO/IEC 27001/Certifications

MINIMUM REQUIREMENTS FOR INSURANCE COMPANIES AND OTHER FINANCIAL INSTITUTIONS
In the non-banking sector, the national requirements also include the minimum requirements for insurance companies (MaGo (Mindestanforderungen an die Geschäftsorganisation von Versicherungsunternehmen) and VAIT (Insurance Supervisory Requirements for IT)) and capital management companies (KAMaRisk and KAIT (Capital Management Supervisory Requirements for IT)) as well as requirements for payment and e-money institutions (ZAIT (Payment Services Supervisory Requirements for IT)).

Basic concepts

ANONYMITY
Data is anonymous if it cannot be related to a specific natural person. Often one encounters the view that this simply means that the individual data records no longer contain a unique key. However, a unique reference can also arise from the uniqueness of a dataset, such as when it consists of a historical time series of payments. This makes it very difficult to anonymize data in many cases and limits the use of anonymized data for training models.

PERFORMANCE
Usually, performance of an AI system or its underlying machine learning model is understood as something like the frequency of correct predictions on a test dataset. It should be noted here that although this measure is optimal in a sense, it is still not an absolute, but a random and potentially variable measure over time, since the test dataset is usually chosen randomly and data can change over time. The performance

of an AI system should therefore also be evaluated in terms of stability and robustness over time. Since transparency and explainability of ML models promote them, the often erroneously propagated contradiction between performance and explainability also vanishes.

PROBABILITY
Too rarely is the central role of the concept of probability in AI applications still addressed. This is all the more problematic because this term is by no means clearly defined, and the simple approach of simply standardizing probabilities in the AI environment instead of classical metrics comes to nothing. Strictly defined, the term means only relative frequency in the case of idealized experiments repeated under exactly the same circumstances. Since machine learning models are set up and used under the same assumptions, this is not conceptually problematic, but the validity of the assumption is usually not given as soon as it comes to modelling human behaviour as is typically done in financial models.

For this reason, the Bayesian notion of probability in terms of subjective expectations, because they depend on a priori assumptions, is much more appropriate for applications in finance. This has far-reaching consequences for the kinds of standards that are established for AI systems based on ML models for Bayesian probabilities.

DELIMITATION OF MODEL RESULT / USE OF RESULTS
When evaluating the use of model results, it is important to conceptually distinguish between the produced output of the model and its use for (possibly automated) decision-making. Often, for example, a probability predicted by a model can be used directly for use in risk management, such as the probability of default in pricing. The correctness of a decision derived from this probability in an individual case is generally irrelevant here.

In other cases, however, such as the automatic rejection of online credit applications, the focus shifts to the decision in the individual case. This includes, in addition to the result of the forecasting model used, a decision algorithm, which is also to be considered a component of the AI system and must meet its own requirements and standards.

Delimiting the financial sector
The financial sector includes all companies that have services in money matters as their core business. This includes, among others, banks or credit institutions, insurance companies, capital management and investment companies, payment service providers or stock exchanges. Financial institutions are supervised by the European Banking Authority and the German Federal Financial Supervisory Authority (BaFin). BaFin is also the authority in Germany that grants official permission to operate financial institutions.

Financial services institutions provide financial services to private customers, corporate customers, public-sector entities or other banks. Such services include: The safekeeping and investment of finances, financing of projects through loans, insurance policies, transactions, securities transactions, etc.

Artificial intelligence is already being used in many processes in the financial sector. These include authentication procedures of new customers, speech analytics and texting, risk analysis, credit scoring in lending processes, AI-based financial investments, and the creation of individualized insurance policies.

4.8.2.2 Knowledge bases/search engines

AI development in the financial context often does not use pure source data. Instead, data must be prepared and typically brought together from structured and unstructured sources, sometimes stored cross-linked for later use in knowledge bases ("knowledge graphs"). This can be done in batch processes, but also ad hoc via search engines. The latter can be publicly offered web search engines or self-developed search systems.

The central component here is the entity or identity matching. There are a wide variety of methods for matching data to the same identity in different datasets.

For the quality of an AI based on linked data, the process of linking data, especially to individuals, is crucial and should meet certain standards if it is not fully deterministic (e.g., via unique technical identifiers). The focus here is on data linkage. The later training of models is not yet considered.

Standards for exact entity matching
In the simplest case, the assignment of identical identities can be done using a technical unique key present on both sides. Apart from data quality issues, this is not critical and can be applied without further control.

In some cases, keys are used that are not technical in nature, but typically consist of a combination of record attributes that are believed to establish a unique reference to the entity, such as surname, first name, date of birth, and place of residence in the case of natural persons; Section 111 of the Gesetz über Ordnungswidrigkeiten (OWiG) (German Law on Administrative Offences), Section 111 False names (see [387]), for example, contains corresponding attributes. These methods are also usually used without further control and are considered accurate, even though they have a certain probability of error.

Depending on the criticality of an AI application, certain standards should be defined for what data sources and attributes can be considered sufficient for accurate matching, distinguishing between natural and legal persons where necessary.

Standards for probabilistic entity matching
Common in the field of AI, especially that based on big data, are matching methods that use machine learning models to find identical entities in different data sources with a certain (high) probability. Since these so-called fuzzy matches are frequently reused, a standardization of their quality is particularly important.

Therefore, standards should apply to which data sources may be used at all for applications of different criticality, and above which criticality inaccurate matches are generally not allowed, for example in high-risk applications such as lending.

In addition, there should be clear rules for specifications on the accuracy of matches and on the subsequent reproducibility of data links that have been corrected, where necessary. Here, a level system for identity matching (unique IDs, high confidence level, medium confidence level, etc.) would be conceivable.

Standards for search engines
Basically, search engines are also based on the principle of inexact entity matching. However, the matching process is ad hoc and the user usually determines which search results are correct matches. Despite this possibility of control by a human user (if one is involved), the quality of the model-based matching is again crucial for the correct assignment of entities. This means that the same standards should apply here as in the case of batch processing.

Additionally, two other challenges arise when using search engines: Since these are used ad hoc, the matching model used is constantly changing, which compromises the reproducibility of the allocations. On the other hand, user activity itself also changes the search result, especially in the case of commercial search engines. Frequent requests for information can affect credit scores in certain cases.

To address the problems of variability of search results and feedback effects of search queries, standards for the reproducibility of search results and transparency of feedback effects should be established. In certain contexts, only search engines or services where the query cannot affect the outcome of future queries may be used.

Standards for influence and control
In addition to standardization for the technical quality and transparency of exact and probabilistic matches, standards for personal control capabilities for the data linked under one's identity could be useful.

This could include standard access points for control, consent to use, and correction of probabilistically linked data. This should include information on the origin of the matches or also the definition of a desired confidence level for matching, for example with search engines.

One way to achieve this would be to use identity standardization systems such as SSI (self sovereign identity) or a bank ID as in the Nordic countries[94]. This could also include standards for information sharing between platforms, including stakeholder participation in portability.

4.8.2.3 Individualization / Fairness

By their economic nature, financial services always relate directly or indirectly to people. Therefore, the challenge also arises for AI applications in the financial sector that the machine learning models on which they are generally based should statistically represent human behaviour. Humans, however, unlike machines, autonomous vehicles, and even to some extent the medically relevant biological processes in the human body, are highly individual, variable, and difficult to predict in their behaviour. This poses particular difficulties for the application of statistical modelling such as machine learning for proposing fair decisions and forecasts, from which specific standardization needs arise. These will be examined in more detail below.

[94] s. also https://www.crefotrust.de/

Fairness

Fairness is frequently seen as an operationalization of non-discrimination, but it often goes beyond that. There are over 20 different basic methods of measuring fairness discussed in the literature; in addition, there are several variations of most of these methods. Many fairness measures follow different philosophies of fairness and are thus in conflict with each other. If the goal is to satisfy multiple fairness philosophies, there are several options. For example, minimum thresholds can be defined for multiple, conflicting measures, or multiple measures can be used in a weighted fashion to calculate an overall fairness score. Which fairness measures are appropriate and which are not under which circumstances has so far only been scientifically established for a few measures. There are known applications where the choice of an inappropriate measure has resulted in societal harm. Most fairness measures are based on quality measures (comparing the quality of two subgroups that differ in only one sensitive attribute in some way). An important aspect of fairness in relation to ML is that the term is always considered in relation to the target criterion of the model. If this is objectively given, fairness should always be considered downstream as a goal of development.

Transparency and explainability

Machine learning models are often referred to as black boxes, which is wrongly understood to mean that the results of such models are not comprehensible in detail or even randomly. Both are usually wrong. The results are mathematically unambiguous formulas, and only the correctness of the prediction is random, not the prediction itself, which is to be calculated deterministically from the input values. A lack of transparency only arises from the possibly complex formulas describing an ML model and the fact that these are not derived from deeper contexts and a logical form, but a general formula was adapted to sample data via free parameters. Therefore, the explanation ex ante by the underlying theory is missing, which can only be established by analyzing the model ex post.

Here it is crucial that explainability must necessarily be based on (at least approximate) causality. An explanation that is not causal and thus correct under all circumstances (i.e. that is not only suitable for justification ex post in individual cases) is at least incomplete and thus does not meet the very claim to an explanation. Therefore, the term explainability should be avoided in favour of the term transparency, and should not appear in any standardization.

Which statistical statements may be applied to the individual?

The discussion of fairness often takes issue with unequal treatment in terms of a crude attribute, such as gender or ethnicity.

However, the question of fairness also arises in reverse at a very elementary level: From which group attributes is it even fair to infer the individual from the group? All statistical, non-causal models, i.e. also all ML models, but also all rule-based models (even in these the correctness of the decision is often a random variable) decide on the basis of input data that are as relevant as possible, but never completely represent the individual (causally). This is inevitable and necessary and should not be questioned.

Here again, the application is of decisive importance. Especially when it comes to (restriction of) fundamental rights, there should be a possibility to look at the individual case. In particular, this means including at least all available relevant individual factors. A current example is the assessment of risk of coronavirus infection, whether in the context of societal constraints or, currently purely hypothetical, such as in the context of insurance rates. This should include the right to include, for example, the detection of antibodies (or other relevant factors) in the decision model or the individual decision. An assessment of the hazard without considering such a relevant factor cannot be considered fair, at least as long as no direct disadvantage to others results.

The problem can often be observed when the groups formed according to the decision criterion differ statistically only slightly with respect to the target criterion, i.e. a large variance can be observed in all groups and the value ranges overlap. For example, although women's and men's incomes are significantly different on average, the distributions overlap substantially. If gender were used to predict salary and thus creditworthiness, it would be unjustifiable for the individual. Technically, this would be noticeable by a high error rate in the prediction of the model and should therefore be noticeable in a good model validation. However, in some cases, users are not aware that even the use of simple averages for decision-making is a predictive model, such as in the risk assessment mentioned above. For credit scores, the formation of customer groups is very relevant. It should be noted here, however, that the latter usually only influence the risk price and not the credit decision and are therefore more justified.

Therefore, to ensure the fair application of models to groups of human individuals, standards are needed that clearly define when to assume complete coverage of all significant influencing factors and, in the case of decision systems, sufficient statistical separation of groups.

Usage note

Completeness of training data
Machine learning models, like all models, should approximate reality in all relevant aspects. They are therefore complete precisely when they cover all relevant contexts. In contrast to scientific models, the basic regularities whose parameters are to be learned from the data are usually unknown for machine learning models, e.g., for problems in AI or for financial applications. They must themselves be learned from the data, which would generally require a continuum of sample data. This means that completeness can only ever be defined pragmatically, but never exactly, for example in the question of which factors and sample data points allow the probability of repayment of a loan to be correctly determined in all conceivable situations.

Non-discrimination in financial decisions
As a counterpart to the problem of unfair equal treatment of members of a group considered in the first subchapter, we look here at the more frequently discussed issue of unfair decisions due to unjustified discrimination, i.e. unequal treatment. The concept of "equal treatment" is problematic and will not be discussed here. Standards for a definition of equal treatment are an important prerequisite for operationalizing fairness.

Preliminary consideration: Fairness in AI applications as a downstream concept
Fairness is always a downstream concept in the context of AI in the financial sector, as the AI application must first be realized for its actual purpose, often a statistically correct risk assessment. Only then can one meaningfully talk about and ensure the fairness of the application. This can be justified as follows.

Fairness as a fuzzy concept
Fairness or justice are not fixed concepts. Typically, this is understood to mean that certain groups are (a) treated "the same" or (b) treated equally taking into account objectively relevant criteria.

Already these two views (a) and (b) contradict each other, and there are many mathematical definitions of fairness that can be shown to be never satisfied simultaneously. Fairness should therefore be considered as a property of an AI system as far downstream as possible.

The task of machine learning and other methods to build data-driven decision systems
Typically, in terms of fairness of a system, we think of AI systems as providing a decision/suggestion regarding a natural person for given input data, along with a probability of the decision being correct, if applicable. If the proposal is based on human-defined rules, these can be directly considered for fairness. More relevant for us are systems where the assignment of input data to a decision is done by a mathematical function (a "machine learning model") that has been determined from sample/training data to produce "best possible" outputs for new, unknown input data (this is where the expertise of the modeller comes in).

The optimal representation of the training data by the ML model usually happens without considering fairness aspects. This is important because any intervention in the learning process would partially corrupt the above-mentioned task of machine learning – if only against the background of the fuzziness of the concept of fairness as described above.

However, there are methods that may improve fairness while maintaining the same predictive quality, provided a concrete fairness measure is given and the model found actually violates it. A differentiation is made between

→ **Preprocessing:** e.g., modulation of the datasets for fairness with the same information content
→ **Inprocessing:** e.g., fairness as a parallel learning target (part of the target function)
→ **Postprocessing:** e.g., outputs with high uncertainty are manipulated to optimize fairness

In the practice of financial applications, however, these are less relevant, since fairness must be explicitly realized.

Practical implementation of fairness
When the AI system is finished training with the ML model included, a review of the extent to which the decisions are "fair" according to a measure to be specified must be performed. If a lack of fairness is observed, this can have several causes.

If errors are found in the model, the model should be checked, especially with regard to its transparency and causality, also with regard to the completeness of the considered features ("further/other columns in the training dataset").

If a lack of fairness is already evident in the training data and if the model is valid, there appear to be objective causes for the inequality. Maybe the training data are not representative for all groups, then one can try to supplement them ("further/other rows in the training dataset"). If this is not the case, a definitely objective but unintended inequality can be compensated ex post by averaging the results over the relevant group attribute. For this, the corresponding attribute, e.g. gender, must be known and explicitly included in the model.

In addition, there are cases where no objective target criterion was used in the training dataset (e.g., true loan defaults), but human decisions that were already biased. This may also arise implicitly from the fact that, for example, certain loan applications are not accepted from the outset and therefore their default behaviour cannot be observed. Creating usable training data here is very difficult.

All these measures basically do not interfere with the modelling process, but concern either the provision of the data or the ex-post treatment of the results.

Another important aspect arises from the fact that fairness cannot be discussed separately, but is interdependent with the issues of transparency/explainability and completeness of training data. In particular, there is no contradiction or "trade-off" between fairness and performance or fairness and transparency, any more than there is between fairness, transparency and performance. Performance and transparency as well as performance and selection of training data are interdependent and have to be optimized together.

The assessment of fairness, on the other hand, is causally dependent on performance, transparency, and correct data selection. However, from the observation of an unfairness of decisions as described above, it can be concluded that there may be deficiencies here. This does not have to be the case, however; there should be no direct intervention in model building just to achieve fairness.

For standardization, this results in the requirement to define fairness measures, but, at least in the context of financial applications with risk relevance, not to impose any requirements that would, for example, jeopardize the unbiased representation of the observed data in favour of fairness a priori.

Fairness concepts
Why is fairness of particular relevance in financial decisions?

LEGAL CONCEPT OF FAIRNESS
Particularly in the U.S., the different legal fairness concepts "disparate treatment" and "disparate impact" have become established for the insurance industry. It is necessary to define what constitutes proxy discrimination, i.e., (un)intentional discrimination by a proxy variable, such as a postal code instead of ethnicity, where there is a causal link or high correlation with the prohibited characteristic. In addition, it is necessary to define what is meant by risk-adequate differentiation and what constitutes discrimination, see [388].

MATHEMATICAL CONCEPT OF FAIRNESS
In addition to legal concepts of fairness, there are also mathematical concepts of fairness. An overview of these is given in [389]. There, a distinction is made between individual and group fairness measures (see list below).

Selection criteria must be established for the use of a mathematical fairness measure. Tolerances must also be specified, as these dimensions generally cannot be met exactly. It must also be made clear that multiple measures of fairness generally cannot be satisfied simultaneously. For example, independence, separation, and sufficiency can only be satisfied if the data are already "fair" in themselves (see [390], Chapter 2, for example Proposition 2, for such statements).

FAIRNESS OF DATA
Data as such can also be unfair. For example, certain groups may be over- or under-represented. Also, existing disadvantages may persist through use of data.

There are approaches, called "debiasing techniques", to prepare data before actual machine learning in such a way that the bias is removed. Examples are disparate impact removers or orthogonal predictors (see [389]).

DEFINITION OF FEATURES WORTHY OF PROTECTION
How are features worthy of protection defined? How should borderline cases such as postal code, education level, score for creditworthiness be dealt with (see also [388])?

There are more than 20 measures based on either one or more quality measures (distributive fairness, group fairness) or on a distance measure (individual fairness). In addition, for most fairness measures there are a number of variations, fairness measures can be combined with each other, and it is possible to develop/define one's own fairness measures.

Conflicts between fairness measures
When choosing a fairness operationalization, there is usually no definite right (only possibly a definite wrong). Different specific applications as well as different stakeholders can influence an appropriate choice, but different stakeholders can also have different goals and thus prefer different measures. In this respect, similar questions as well as approaches to solutions are offered in the choice of an explanation.

Different measures of fairness represent different notions of fairness, and many of them cannot be optimized at the same time because they conflict with each other to some degree. If optimization is targeted for a particular fairness measure, this will inevitably reduce the results of other fairness measures. This can actually increase discrimination according to the understanding of reduced measures. In addition, it is possible to specify targeted fairness measures that achieve a very high value but do not necessarily coincide with specific fairness perceptions. Therefore, it makes sense to justify the choice of fairness measures if the resulting values are communicated or made transparent. If an external party (e.g. an audit authority) wants to check whether the transparently stated values are actually achieved, the implementation of traceability mechanisms such as assurance cases is a good idea.

Processes for choosing appropriate fairness measures include design thinking, specification workshop, assurance cafés, or stakeholder engagement. The choice potentially falls on competing fairness measures that cannot be optimized simultaneously. There are several ways to deal with this conflict situation:

→ Weighting: An "importance" is defined for each measure, e.g. in the form of a factor (this is called weighting). If the weightings add up to 1, they correspond to a percentage value that the respective measure contributes to the overall score. An (argued) minimum value is set for this overall assessment. If this value is reached or exceeded, the system is considered fair.

→ Thresholds: An (argued) minimum value is set for each measure. If the specified value is met or exceeded for each measure, the system is considered fair.

Embedding in the social process
The choice of fairness measures and conflict resolution strategies can be made arbitrarily, and in the worst case, even purposefully so that a system achieves the best possible scores, regardless of how meaningful the chosen metrics actually are. This means that unfair (discriminatory) "behaviour" is not only not detected, but even deliberately concealed. This makes it all the more important, as already explained, to make sound decisions when making a choice, as well as to document the reasoning. The added value is not only a good basis for argumentation in case of an audit (e.g. certification or litigation), but also the option to be able to justify oneself to a customer or affected party (e.g. as a CDR measure but also for marketing reasons)

Communicating fairness measures to non-experts can be difficult, as without specific prior knowledge, a good basic understanding of mathematics is necessary. It is therefore not sufficient to state the choice of fairness measures and the values achieved in each case, but these must also be explained in a way that is appropriate to the customer or stakeholder. For many fundamental measures, there is sufficient scientific basis to be able to anchor needs-based explanations in standards.

Beyond the choice of measurements and the values achieved, how this information is used also matters. It can be used as (additional) target functions in the context of an AI training process, as minimum requirements (requirement engineering) in the context of a quality assurance process, or even as control specifications in the context of regular automatic checks in the field. The importance of communicating about the use of fairness measures can be illustrated by a simple example: As explained, most fairness measures are based on a ground truth. However, this ground truth does not necessarily correspond to the real-world circumstances in which an AI system is deployed. This means that even if fairness measures achieve sufficiently good values under laboratory conditions based on test data, this does not necessarily say anything about fairness performance in real-world applications. If the selected fairness measures are now automatically calculated on a regular basis in real operations, and if a failure is reported in case of non-compliance, a much more effective quality assurance process is created.

Overview of fairness measures
Individual fairness measures

→ Fairness by awareness [391]
→ Fairness through awareness [393]
→ Counterfactual fairness [394]
→ Controlling for the protected variable

Group fairness measures

- → Demographic parity (or statistical parity) (independence)
- → Disparate impact (the four-fifths rule)
- → Conditional demographic parity
- → Separation
- → Sufficiency

4.8.2.4 Information security

IT security[95] as an important horizontal building block was already a prominent topic in the first edition of the Standardization Roadmap AI. Despite the vertical area of financial services and the sub-area of information security[96] being addressed for the first time in this 2nd edition of the Standardization Roadmap AI, all the basic requirements continue to apply here as well. In this context, the goal of IT security, with its reference to the use of information technology and its intersection with information security, continues to be maximum protection against operating errors, technical failure, catastrophic failures and deliberate attempts at manipulation ([63], p. 99). The technical analysis should also include unintentional manipulation as a part of faulty operation, as well as damage in addition to catastrophe-related failures. The protection targets of information security, confidentiality, integrity, availability and authenticity are also the basis for further considerations here, see Figure 45. The aspect of security is in the foreground. Data quality, which is particularly important in the financial industry, continues to be essential for the successful use of AI, but also for the consideration of information security. General data security measures and confidence levels regarding the data quality of input data are already described extensively in the 1st edition of the Roadmap. Furthermore, it is important to define mechanisms that make a statement about the data quality as well as the possible uses related to the quality of these data. The use of the data is to be designed in a risk-oriented manner depending on the quality of the data. High-risk business processes should therefore be subject to stricter data quality requirements than less risky business processes.

The risk assessment for IT security, attacks and a selection of defense mechanisms are also addressed for the general case. Furthermore, the 1st ed. of the Roadmap contains research results on laws, standards and specifications as of 2020, which have been brought up to date in this 2nd ed. of the Roadmap.

Trust is the foundation of any kind of business relationship. Therefore, in general, but in the financial industry in particular, there is the challenge of establishing trust in the IT security and information security of the provider as well as in the AI system. For this reason, on the one hand, it is necessary to ensure verifiability, explainability, and proof of compliance (as required, for example, in the AI Act draft for high-risk systems), as already stated in the Roadmap, 1st ed. On the other hand, all relevant stakeholders should be included in the individual risk assessment of information security at the very beginning of the development process of an AI according to how they are affected. Stakeholders include, in particular, a company's board of directors, the information risk officer, and risk management. In each financial institution, there are still direct persons responsible for AI systems and data, who identify, monitor and mitigate the risks for the institutions themselves, but in particular also for the customers. Here, a special focus should be placed on vulnerable consumer groups, insofar as they play a direct or indirect role in the respective AI system under consideration.[97]

[95] IT security refers to a state in which the risks present in the use of information technology due to threats and vulnerabilities are reduced to a tolerable level through appropriate measures. IT security is therefore the state in which the confidentiality, integrity and availability of information and information technology are protected by appropriate measures. [BSI (2022)]

[96] Information security has the protection of information as its goal. This information can be stored on paper, in computers or even in heads. The protection targets or basic values of information security are confidentiality, integrity, availability and authenticity. Many users include other fundamental values in their considerations. [BSI (2022)]

[97] Proposed definition: A consumer in the sense of the BSI's Digital Consumer Protection is any natural person who incurs or could incur an IT security risk during the private use of products, services or applications. [BSI (2021c)]

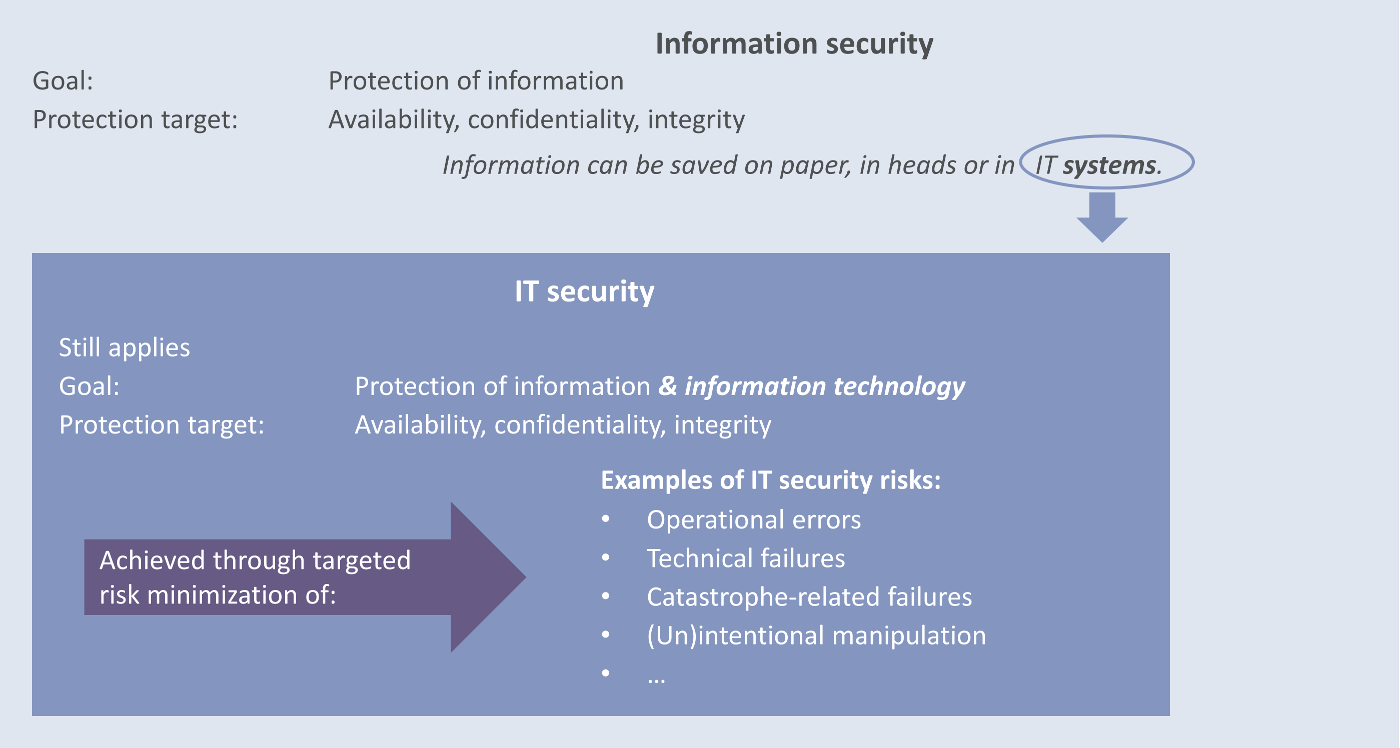

Information security
Goal: Protection of information
Protection target: Availability, confidentiality, integrity
Information can be saved on paper, in heads or in IT systems.
IT security
Still applies
Goal: Protection of information & information technology
Protection target: Availability, confidentiality, integrity
Achieved through targeted risk minimization of:
Examples of IT security risks:
• Operational errors
• Technical failures
• Catastrophe-related failures
• (Un)intentional manipulation
• ...

Figure 45: Information security (Source: Working Group Financial Services)

Special requirements for information security

One of the key elements of business models in the financial industry is trust – trust in the organizational and material performance of a financial institution, as well as the security (including all protection targets) of the data. As a rule, financial and insurance products are abstract, not directly tangible and difficult for non-experts to comprehend in their complexity. Particularly sensitive and, to a large extent, personal (financial) data is processed. This results in the special situation that customers have to trust their financial institutions in several ways, for example, in taking into account individual life circumstances when making individual investment decisions. Financial information in the wrong hands and/or incorrect credit decisions can have existentially threatening consequences for both customers and financial institutions. Financial service providers therefore have a special responsibility. It must be possible to trust that the service provider promise will be fulfilled, that the IT and AI systems behind it are functioning properly, and that appropriate information security measures have been put in place. For this, it is important that the financial service provider can recognize an erroneous result and then also distinguish whether this is rooted in the model (risk) or has occurred due to a successful attack on the AI.

Information security measures must not only lead to greater security in theory, but must also be implemented in such a way that they are manageable for bank employees or customers. This applies to the use of (security) technologies as well as security requirements (management requirements), so that they are actually used as intended and are not omitted, bypassed or used incorrectly.

Usable security in the sense of taking the right and necessary measures is achieved by creating transparency, usability, accessibility and acceptance. The aim is to develop AI systems from the user's perspective. This includes the fact that the respective employee or customer works with a comprehensible user interface, is sufficiently informed or trained, and that security processes run without user intervention. Usage errors that can compromise security are thus minimized. Employees of the financial institution must be adequately trained in the use of the AI system so that they understand how the application works and how it fits into the overall process. Consumers, however, are generally unfamiliar with the internal processes of financial institutions and must therefore be adequately informed about the use of the AI system. Thus, usable security leads to an appropriate level of security, but also to a greater efficiency and performance of the systems.

AI-specific challenges

For the financial sector, there are a number of regulatory requirements for IT that are mandatory for financial institutions. In addition, there are current standards and specifications to be applied in the industry. Although these are not binding, the regulatory framework (see Chapter 4.8.2.1) requires that they be aligned with common standards. Accordingly, these requirements also apply to AI. The regulatory requirements oblige financial service providers to establish a risk-adequate internal control system (ICS) as well as an "information security management system" (ISMS) and to prove their functionality, also for the use of corresponding insurance services. In addition, the IT-related measures must correspond to the state of the art. Attacks on AI systems are not specific to the financial services sector, but there are already increased information security requirements for financial services providers due to the high need for data protection. Crises in the financial sector or even just a loss of confidence in institutions can have far-reaching consequences for the entire economy. The macroeconomic consequences can be greater than in other industries. The decisions made by AI systems of financial service providers affect customers (e.g. lending) and also other stakeholders. So if one classifies the consequences of security problems, then all these groups of people must be included.

The information security requirements resulting from current standards and specifications for IT systems must also be implemented in the technical processes In IT, a suitable combination of static and dynamic analyses is required here in particular to detect and eliminate potential security gaps and vulnerabilities at an early stage (before they are exploited). These analyses include vulnerability scans of third-party libraries (including open source), various code analyses and penetration tests. The same applies to safeguarding the infrastructure, organizational procedures, processes, etc. No special need for standardization is seen here; existing standards implicitly include AI systems.

However, in the case of AI systems, additional attack types and attack scenarios are possible and need to be addressed specifically. The document "Secure, Robust and Traceable Use of AI", published by the German Federal Office for Information Security (BSI) [83], names, among others, evasion/adversarial attacks, data poisoning attacks, privacy attacks and model stealing attacks as AI-specific attack types. Evasion/adversarial attacks and model stealing attacks have already been dealt with in the first edition of the Standardization Roadmap AI (cf. [1], p. 109 ff.); there is no additional need for standardization here. In addition, there are data poisoning attacks and privacy attacks. The BSI defines these as follows:

Data Poisoning Attacks

By manipulating the training data of the AI model, attackers cause it not to respond to (certain) inputs as intended by the developer. Due to the large amount of data and the lack of transparency, these attacks are usually difficult to detect.

Privacy Attack

Attackers extract information regarding training data from the model. Model inversion attacks extract training data and membership inference attacks determine if a datum was used for training

Thus, there is a need for standards for suitable measures that adequately mitigate these attack scenarios. Regulatory requirements and common standards that exist for the use of IT in general must be considered against the background of a potentially changed risk situation. Based on this, additional protective measures need to be considered that target the specific threat situation of the use of AI and are then integrated into the existing risk management system.

Productive data (anonymized as much as possible and reasonable) is often used for the training and validation of models used in AI systems. Thus, these are not to be equated (and especially not to be treated) with, for example, synthetic test data used for the quality assurance of IT systems Therefore, the same protection requirements apply to this data, which is used during the development of the AI system, as to the data that is later used in productive operation.

There is a need for standardization to permit the use of productive data for training and validation purposes, and to take sufficient account of the high need for protection by means of suitable measures. There is a link here to the topic of "data governance", which is addressed below.

Protective measures (information protection)

In the course of considering possible attack vectors on information security, it became apparent that a great deal of attention must be paid to these vectors, especially for financial service providers. This is not so much because there are a greater number of possible attack vectors compared to non-financial service providers, but rather because the protection needs of the data used in the context of AI are generally at least "high" or "very high" out of four protection needs classes (low, medium, high, very high). This is mainly due to the fact that AI use cases in the financial industry are mostly characterized by a short distance to the end customer and, in addition, sensitive and confidential information such as credit rating and health data is processed. Accordingly, it can be assumed that the need for protection of most data when using AI in the financial sector is at least high with regard to the protection targets of confidentiality, integrity and also availability, and thus protective measures must be taken to meet this need for protection.

This increased need for data protection, especially in the financial sector, has already been addressed in the form of existing and comprehensive regulation. Therefore, there is currently no need to comprehensively expand the regulatory framework already in place. Rather, a targeted specification or concretization and reference to these regulations is usually sufficient to adequately address the specific characteristics of an AI during development and productive operation. The technology-neutral approach of BaFin's current supervisory practice, based on MaRisk and its concretizations with regard to IT BAIT, VAIT (insurance supervisory requirements for IT), ZAIT (payment services supervisory requirements for IT) and KAIT (capital management supervisory requirements for IT), generally enables a risk-adequate approach to IT in the context of AI.

In addition to existing regulatory requirements, there are commonly used standards that are applied when implementing information security and information risk management systems. The most important standards include the DIN EN ISO/ IEC 27000 [479] series and the "BSI-Grundschutz" of the German Federal Office for Information Security. In addition to these national requirements and standards for IT, including AI, there are other specific initiatives at the EU level regarding AI and data: e.g., the EU AI Regulation (Artificial Intelligence Act, see Chapter 1.4). This regulation is intended to provide a cross-industry and EU-wide legal framework for the development and use of AI. Furthermore, the EU Commission has set two initiatives in motion as part of its Data Strategy: the Data Governance Act, which came into force this year, and the Data Act, for which a draft bill is currently being discussed in Parliament. Both legislative initiatives aim to increase data sharing in the EU, make it secure, and to promote innovation through the increased use of data. In summary, a meaningful alignment of existing requirements in the context of information security with the technology of AI is needed.

Therefore, the following is not an isolated comparison of individual attack vectors and possible corresponding protective measures, as was already done in the Standardization Roadmap 1st edition (see: [63], 4.4.2.2 Attack vectors and defense mechanisms), but rather the idea of the holistic concept of the "robust AI platform" is to be initiated. According to this "robust AI platform" concept, data governance and information security measures are integrated from the outset so that information is adequately protected throughout the AI life cycle. An AI platform is composed of system components or their subcomponents that provide the AI. In addition, there are the associated data, processes, and security measures that are found in all life cycle phases of the AI application (see also [63], 4.2.2.2 Challenge 1: "Definition of protection targets on the level of processes and data within the AI component"). The term "robust" refers to existing regulation as well as individual measures in the respective financial institution. Therefore, an AI platform is only robust if it is demonstrably in compliance with existing requirements and laws and addresses the specific challenges of a financial institution that address corresponding information risks.

The target picture should therefore be a generally accepted definition of a catalogue of minimum requirements and "best practice" procedures for an AI platform per component, which is done by referencing and giving detail to already existing regulations and laws. Since it is often not yet possible to fall back on established blueprints and procedures when implementing AI use cases in practice, this is a research field of high relevance. The goal is to accelerate the implementation of AI in Germany, while ensuring that it is compliant, secure and reliable.

4.8.2.5 Risk management

Financial services firms are professional risk managers and have long used AI systems to accomplish this task. They also have a long tradition of managing model risk and are competently monitored in this regard by financial regulators using well-established standards. Therefore, embedding AI standardization into existing risk management and audit processes is paramount for the sector. Relevant aspects of this will be discussed below.

Individual model risk

UNDERSTANDING THE RISKS OF USING MODELS AND THE ASSOCIATED CHALLENGES OF MACHINE LEARNING
AI/ML models share most risks with traditional models; however, these risks are more difficult to identify and assess.

Above all, the quality of the data has a significant impact on the performance of AI/ML models and can be considered as the most important limiting factor for AI/ML.

AI/ML models and conventional models differ in algorithmic risk characteristics such as explainability, bias, and robustness.

Not only algorithmic risks need to be considered, but also legal risks related to data protection law, civil law (for example, responsibility in declarations of intent by AI), anti-discrimination law.

To effectively reduce model risk, it is critical that the model risk is appropriately embedded in all three lines of defense. For this, employees must have the appropriate skills and responsibilities must be clearly assigned.

HOW CAN AI BE INTEGRATED INTO EXISTING MODEL RISK MANAGEMENT FRAMEWORKS?
(Regulated) financial firms need to establish a governance structure that addresses the risks of AI/ML models – ideally based on existing governance frameworks. Governance – which applies to the entire bank – should cover the general model risk management of all productive models and extend it to include AI-specific guidelines.

The first step towards an AI-specific model risk framework is to revise the processes/regulations already in place. The overall approach is no different from the model risk of traditional models, namely from the development of a risk strategy with appropriate risk bearing capacity and adequate risk appetite to risk mitigation measures.

→ Development and validation considerations should be driven by conceptual soundness, data and feature engineering, training and calibration, and testing and monitoring. In this context, finding the most appropriate model and its parameters, as well as risk assessment, model change management, ongoing modification, problem management, the software development process, and supplier model management should be considered.
→ In addition, model owners as well as model developers must make an assessment of whether an AI/ML model will provide the desired performance improvement or whether a traditional model is sufficient to meet the bank‘s risk appetite in terms of model risk.
→ It should be ensured that the test data are representative, especially with respect to the productive data, in order to be able to measure generalizability. The data quality of the training data should also be guaranteed by ensuring that it is accurate, complete and free of inconsistencies. In the area of supervised learning, it should also be ensured that the annotations of the data are of high quality.
→ Model complexity can be assessed by evaluating three subcategories, namely model input data, assumptions & theory. Here, the implementation should not influence the complexity of a model, but the number of free parameters, although these are not always obvious when, for example, complex features are generated.
→ In general, models should be assigned to different complexity classes with corresponding due diligence requirements. An exact standardization of this seems unrealistic for reasons of complexity, therefore a rough qualitative division into a few classes should be made.
→ A continuous, largely automated monitoring is to be defined, replacing the previous annual validation cycles with event-driven reviews triggered by the automated monitoring.

Like conventional models, AI/ML models must undergo initial and periodic review and validation, depending on the level of model risk

How can AI be integrated into the existing model validation framework?
AI/ML algorithms bring special challenges, such as explainability and robustness, which need to be implemented in a regulatory internal ratings-based approach (IRBA) model, especially for banks. The complete data history (data lineage), the model parameters and generally all metadata must be accessible at a later time. The model validation process needs to be extended to address the specifics of AI/ML. As with traditional models, validators must ensure that the selected model is conceptually sound by checking that the individual features are generatable predictors and that they make sense from a business perspective, and that the model is not overfitted to irrelevant aspects of the training data. The process of model validation needs to be improved to meet the specifics of ML. Model validation must ensure that the model predicts as intended in unforeseen situations, for example by performing stress tests. Model validation must also ensure that the model is sufficiently transparent to the various stakeholders, i.e., that it is able to explain the rationale for a particular decision. Ways to plausibilize model results a posteriori include explainable-AI methods such as: Perturbation analysis, gradient analysis, surrogate modelling, and example-based explanations. Bias within the data must be considered in the design phase – starting with the selection of input variables.

Although this is largely done automatically in the course of model optimization, bias can only be detected at all with respect to available input variables, for example.

Risk modelling / correlation model for measurement of cumulative errors
For successful risk modelling in machine learning/artificial intelligence, it is important to consider the particular risks of such a system. Often ML models are interlinked or applied to entire portfolios, potentiating possible risks – as with conventional models, but the implications and tools available to respond in the event of failure differ. The following possible risks and risk dimensions are to be considered here.

Possible risks and risk dimensions

OUTPUTS FROM CORRECTLY FUNCTIONING SINGLE MODELS ARE USED INCORRECTLY IN OTHER SYSTEMS
If results of individual models are used purely on the basis of their classification, this can lead to unexpected system behaviour. For example, if the classification decision proposed by one model is reused in other models, such as a rating class rather than the exact credit score, this introduces bias even in a simple averaging exercise. For the use of model outputs, the original context should be known and taken into account. This includes: Model decisions and weighting, the origin and context of training data, the origin and context of real-time input data. Standardization can start here and ensure that no relevant information remains unknown during the acquisition process.

CONTEXT-SENSITIVE MODELS ARE USED IN OTHER ENVIRONMENTS / LINKED FROM DIFFERENT ENVIRONMENTS
In contrast to classical models, decisions made by artificial intelligence are not easy to understand. If models are fed to a new context or linked from different contexts, this can lead to fundamentally wrong decisions. For the use of other model outputs, the original context should be known and taken into account. This includes: Model decisions and weighting, the origin and context of training data, the origin and context of real-time input data.

A possible cost-saving through reuse should always be considered in a differentiated manner, since an apparently good model can deliver fundamentally different results in a new context.

Standardization can be useful here and help to ensure that essential information is used.

ACCUMULATION OF DEVIATION ERRORS
Some input values change over time. Manual reweighting of parameters is usually impractical in machine learning. To adapt the model, adjustable parameters can be integrated from the beginning to account for such developments – another option is regular re-evaluation based on new training data and the requirements catalogue defined at the beginning of the project. Since input values vary more and more over time, the effort increases the longer a model has not been re-evaluated. To reduce hurdles, standardization could create conditions for regular and low-threshold re-evaluation.

Placement in the existing regulatory framework and ongoing consultations with the German Federal Financial Supervisory Authority (BaFin)
The view of the regulator, namely the European Banking Authority (EBA), as well as the BaFin and the Federal Bank of Germany (Bundesbank), is presented below. Since the consultation phase has not yet been finally concluded, the current discussion papers on this topic are used for this purpose. With regard to the BaFin's MaRisk, it is expected that no further regulations regarding the use of artificial intelligence are currently necessary and that corresponding models are covered by existing rules.

The EBA's paper on the use of machine learning in the internal ratings-based (IRB) context provides a set of principles-based recommendations to ensure that machine learning models in the context of the IRB framework comply with the regulatory requirements set out in the Capital Requirements Regulation (CRR).

According to Art. 179 of the CRR, IRB models must be "intuitive". This means that there must be an easily understandable link between the risk drivers and the default indicator for PD (probability of default) models. Conventional models meet this requirement: They often show very clear and immediately quantifiable relationships between a risk driver (e.g., loan-to-income) and the "yes/no" default.

The EBA recommends that institutions ensure that governing bodies are able to understand assumptions, limitations and the theory of the model by providing them with adequate documentation. In addition, the employees in the model development, credit risk controlling and validation departments must be sufficiently qualified. The building blocks of fairness, explainability, and robustness are particularly relevant for quantitative model validation.

For ML-based models, it is usually more difficult to document the design, the functional details, the theory underlying the model, and the modelling assumptions (see CRR, Art. 175). Therefore, institutions are advised to avoid overly complex modelling decisions unless they are justified by a significant improvement in predictive ability. Above all, bias is to be avoided, so that business decisions must not be based on systematically distorted results that put individual customer groups at a disadvantage. This is necessary, above all, to comply with prohibitions on discrimination in EU legislation and to minimize the resulting reputational risks. This is an overriding principle, as bias can affect both development and application. Also, reference is made to differentiation prohibited by law. For example, certain characteristics such as origin, gender or sexual orientation are not to be taken into account in risk and price calculations.

During the development phase, it is important to consider the following points (cf. IRBA):

- Appropriate data strategy and data governance (including representativeness)
- Compliance with data protection rules
- Correct, reproducible and robust results
- Appropriate documentation
- Appropriate validation processes

During the application phase, it is important to pay attention to the following points:

- “Putting the human in the loop“
- Intensive approval and feedback processes
- Establishment of emergency measures
- Validation, evaluation and adaptation

To ensure that the model is interpreted correctly, institutions are advised to do the following when analyzing models:

- Analyze in a statistical manner the relationship of each risk driver to the output variable and the overall weight of each risk driver in determining the output variable.
- Evaluate the economic relationship of each risk driver with the output variable to ensure that model estimates are plausible and intuitive.
- Create a summary document that explains the model in a simple way based on the analysis results.
- Ensure that potential biases in the model (e.g., overfitting to the training sample) are identified.

In order to bring the use of AI into already recognized structures, the joint discussion paper [395] of the BaFin and the Bundesbank focuses on supplementing, specifying and further developing existing regulations and is intended to serve as guidance for supervised entities. In principle, existing regulations should be observed as a matter of priority. In particular, the focus is on the internal models for capital requirements under pillar 1 and for risk management under pillar 2.

When considering ML, it is always important to look at the entire process and the concrete situation in the institute and not just the algorithm alone. The appropriateness of an algorithm then depends accordingly on the concrete application or decision-making process as well as on the scope and quality of the data (see also Chapter 4.8.2.4).

Due to the diversity of ML approaches, there is also no universally valid acceptance process. It is necessary to conduct a risk-oriented examination and objection of algorithm-based decision-making processes. Exceptions to this are justified cases, such as internal models, with a focus on methodology, calibration or validation.

Important principles for oversight remain risk orientation, proportionality and technology neutrality. This is another reason why this requires more intensive monitoring when algorithms are used in critical decision-making processes. The same applies when considering complexity, recalibration frequency and degree of automation.

Responsibility clearly remains with the management in this new respect as well. It provides strategies and guidelines for the use of algorithm-based decision-making processes, whereby potentials as well as limits and risks of such processes must be taken into account. Appropriate technical understanding and communication appropriate to the addressee are required for this. Institutions are recommended to create an overarching framework that includes an inventory of all algorithm-based decision processes (model inventory) and considers their interdependence. In addition, it is suggested that this aspect be considered in the model risk management framework.

The BaFin and the Bundesbank currently see no need for a fundamentally new oversight practice for ML methods. However, the extent to which adjustments are required at certain points is being examined on an ongoing basis.

The oversight practice, shaped in particular by MaRisk and xAIT, has endured. The comprehensive rules in pillar 1 and the principle-based requirements in pillar 2 provide a solid foundation. The oversight focus is on new or more pronounced risks in the data basis, validation, model change, and governance. A challenge for oversight is the consistency with the AI Regulation and consumer protection. Three key points for bank oversight are:

1. Models invite data credulity, creating the danger of "overfitting." Presumed correlations based on random properties are identified. As a result, ensuring data quality is a central task of the institutions that are subject to oversight.
2. The focus shifts to the explainability of the models instead of the comprehensibility. In this context, the banking regulator considers explainable AI to be promising, but it must be taken into account that this term also hides models.
3. Material model changes are harder to detect (adaptivity). The line between model maintenance and model changes is blurred. In particular, there is a risk that models may move away from the original model within a short period of time without the model owner noticing.

From today's perspective, the proposals appear to be a good way to extend the existing framework for regulating model risk for the specifics of methods relevant to AI.

4.8.3 Standardization needs

Need 08-01: Definition of verifiable anti-discrimination metrics to demonstrate that an AI solution is non-discriminatory
AI should have as positive an effect as possible, but it must also be subject to rules. Where people are involved in the financial services sector, an important rule is the prohibition of discrimination. The compliance of providers with the rules, the verification by control authorities and the presentation to consumers is a major challenge, partly because the term discrimination is ambiguous and related to other terms such as fairness, justice and equal treatment.

In the following, discrimination is understood as unjustified disadvantage or preference. (Within the meaning of Art. 3 para. 3 of the German Grundgesetz (GG) (Basic Law).) A verifiable – in the best case by automation – definition of discrimination can result from the standardization of metrics in this respect.

Here there are some difficulties:

→ In current research, anti-discrimination metrics are often referred to as "fairness metrics."
→ Moreover, there is more than one discrimination metric in the current discussion.
→ Not all previously known anti-discrimination measures can be complied with simultaneously.
→ Developers of AI solutions must therefore have the possibility of choice.
→ There must be permitted tolerances in metrics compliance when metrics cannot be complied with exactly in practice.

A trustworthy AI can be an opportunity for Europe and European companies to compete with U.S. and Chinese AI providers. The financial sector would particularly benefit from measures, as there are fewer trusted individuals than, for example, with physicians in the medical sector. A "seal of approval" analogous to the "Blue Angel" or the Nutri-Score and/or a rating in the "S" part of Environmental Social Governance (ESG scores) for companies would be conceivable.

Providers and developers of AI solutions benefit from the legal certainty provided by objective and automatically verifiable rules.

Need 08-02: Standardization of the characteristics relevant to non-discrimination and how to deal with them
Anti-discrimination laws and provisions inconsistently identify the relevant characteristics.

Examples:

Charter of Fundamental Rights of the EU (Art. 21 "Non-discrimination"): "Any discrimination based on any ground such as sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation shall be prohibited."

Treaty on the Functioning of the EU: "... discrimination based on sex, racial or ethnic origin, religion or belief, disability, age or sexual orientation."

A unified and conclusive list of characteristics can help to avoid efforts in the creation of AI solutions or improve the performance of an AI solution.

In addition, how the relevant characteristics are to be taken into account when creating AI solutions should be standardized. A general exclusion may be counterproductive. Example: Assuming that a person's creditworthiness depends on the duration of previous bank relationships and that, at the same time, for historical reasons older women in particular have shorter bank relationships on average, a given duration of a bank relationship would possibly be evaluated more positively for a woman than for a man. If the characteristic "gender" were removed from the learning data, older women would be systematically disadvantaged.

Providers and developers of AI solutions benefit from legal certainty through consistent rules.

Need 08-03: Standardization of the consideration of non-discrimination issues in the creation of an AI solution to demonstrate non-discrimination
Another way of proving that an AI solution is non-discriminatory is not to standardize the product/service itself, but to standardize the process of creating the product/service with regard to the consideration of the prohibition of discrimination. The metrics demanded in Need 08-01 can be introduced in that a standardized process requires the use of standardized metrics. In doing so, it must be possible to determine the impact of metrics compliance on the overall performance of the AI solution.

Need 08-04: Definition of the concept of fairness through verifiable metrics
Fairness is a term that is even less defined than discrimination. Unlike discrimination, it is not regulated by law and does not appear in the Charter of Fundamental Rights of the EU at all and appears in the Treaty on the Functioning of the EU only in the context of sport. All the more reason why standards analogous to those mentioned in Need 08-01 and Need 08-03 with respect to "non-discrimination" are needed.

A "fair" AI – i.e., voluntary adherence to fairness metrics – can also be an argument for confidence or a selling point for AI solutions in finance, along the lines of the 08-01 rationale.

Need 08-05: Rules for demonstrating coverage of all relevant factors in group considerations
When AI systems make statements about groups, they are not necessarily transferable to the individual. It must therefore be ensured that either no significant individual factors are missing from the model or that it is possible in principle to assert and take them into account, provided that this does not contradict ethical principles. This is especially true when fundamental rights are restricted based on models that make statements about groups of individuals.

In the context of financial applications, but also in other socioeconomic systems, the focus is often on a risk consideration across the group, such as predicting the expected loss in a loan portfolio or the expected spread of a disease. However, a correct prediction for the portfolio and corresponding risk prices (or, analogously, corresponding health protection measures), must also be optimized for the individual (whose fundamental rights are affected) among all the information available to them. This means that, depending on the severity of the consequences, all individual factors that have been shown to have a significant impact on prognosis must be taken into account. Therefore, rules are needed according to which the relevant factors are determined.

Need 08-06: Development and definition of (minimum) requirements for an AI platform
Guidelines are needed to design a robust AI platform from an information security perspective. This affects not only the purely technical aspects of a corresponding IT platform, but also the procedural design of the development and subsequent operationalization of the AI system. The term AI platform is defined here as the sum of the system components or their subcomponents that provide the AI, as well as the associated data and processes that are applied across the life cycle phases of the AI.

Not only the subsequent operation, but also the development itself places high demands on the information security of an AI platform. In addition to the direct and extended protection goals of information security, the minimum requirements for AI platforms must also take into account data protection requirements in particular.

The criticality of these aspects is particularly high in the area of financial services. This is mainly due to the fact that AI use cases in the financial industry are usually characterized by a significantly lower distance to the end customer and thus information such as creditworthiness and health data is used, which requires increased sensitivity.

Therefore, strict requirements should be placed on an AI platform in the financial services sector. These must be in line with existing specific regulations (BAIT, VAIT, KAIT, etc.). However, in order to meet the increased need for protection, it is not necessary to comprehensively expand the already existing regulatory framework. Rather, a targeted, practical specification and a reference in the form of guidelines and concrete specifications (in the sense of best practices) should be made. It should be noted that future changes or additions to the regulatory requirements cannot be ruled out. This is particularly true in view of the international regulatory projects currently underway.

Need 08-07: Framework for handling training data for AI models

There are extensive (behavioural) requirements for data used for testing purposes in the financial industry. With regard to training data for AI systems, the existing restrictions need to be reviewed in terms of practicability and maintenance of a high need for protection.

For training the models used in AI systems, data from the production environment (anonymized as much as possible and as reasonable) is often used. This means that the training data cannot be equated with (and above all cannot be treated in the same way as), for example, synthetic test data used for the quality assurance of IT systems.

Synthetic test data have no relation to real data and thus do not allow any conclusions to be drawn about such data. The protection requirements for synthetic test data are therefore generally low, and the requirements for handling them are correspondingly so. Rather, here there are specifications that real data may not be used for tests.

However, training data for AI models must allow inference (to a certain extent) in order for the models trained on them to be valid. Thus, their need for protection is significantly greater than that of synthetic test data. Therefore, the (low) constraints for synthetic test data are not transferable to training data; here more extensive regulations are needed.

The training, validation and test data of the AI models thus have the same need for protection as the productive data In the case of financial services, there is usually at least a high need for protection (the highest protection need class for personal data). Here, suitable framework conditions must be created, especially with regard to information security and data protection, which on the one hand take into account the high need for protection and on the other hand allow the training of AI models.

Need 08-08: AI-specific attack scenarios and protective measures

AI is creating a new risk situation in the financial industry, not only by changing the intensity of existing risks, but also by creating new attack vectors. The changed framework conditions must be taken into account in a standardization process. The use of AI in IT systems makes additional types of attacks and attack scenarios possible – from an information security perspective. In order to adequately reduce the risk of such attacks, they must be taken into account as part of information security measures. The document “Sicherer, robuster und nachvollziehbarer Einsatz von KI” (Secure, Robust and Traceable Use of AI), published by the BSI [83], names evasion/ adversarial attacks, data poisoning attacks, privacy attacks, model stealing attacks, among others.

Current standards and specifications for IT systems (without a specific focus on whether AI is used) do not specifically address these attack scenarios or corresponding measures. However, a standard for AI systems should address this.

This need is expressed in the context of financial services, since the security requirements in terms of confidentiality, availability and integrity are (at least) high. This is also reflected in existing requirements and standards for general IT systems through the regulatory requirements of banking oversight. Regulations that are for the use of IT systems (without artificial intelligence) must be considered against the background of a potentially changed risk situation. Based on this, additional protective measures are to be implemented that target the specific threat situation.

Need 08-09: Definition of criteria that are sufficient for automatic entity matching

For critical systems, identities in two different datasets may only be matched if they match 100%. It is therefore necessary to determine which criteria are sufficient for this purpose. Even for non-critical systems, it serves quality if data is assigned to the correct identities.

Example: The customer number cannot be clearly assigned to the person. In the financial sector, the datasets used to train an AI are not always mapped via unique identifiers, such as the personal ID card number or the health insurance number in the healthcare sector.

Need 08-10: Establishment of criteria on how to measure the reliability of matching using static models and what minimum values are necessary

If identities are only matched probabilistically, it must be possible to measure how reliable the matching is and for which type of application which minimum reliabilities should apply.

The incorrect mapping of data to entities is as much a source of error for training and application of AI as the incorrectness of correctly mapped data.

Need 08-11: Establishment of mechanisms for users to monitor the use of their own identity
Users should be able to know what data has been aggregated under their identity. This is already mandatory under the General Data Protection Regulation (GDPR), but it is unclear whether this includes all data that has been added through fuzzy matching, such as magazine articles.

The process of fuzzy allocation is hardly completely safe to monitor. Stakeholder participation would help to significantly improve the quality of the assignment. This has always been a major challenge in the financial sector.

Need 08-12: Guide to usable security
Measures in information security must not only lead to greater security in theory, but must also be practically manageable/implementable from the user's point of view. This applies to the use of (security) technologies as well as security requirements (management requirements), so that they are actually used as intended and are not omitted, bypassed or used incorrectly.

Usable security in the broadest sense is achieved by creating transparency, usability, accessibility, and acceptance. Usage errors that could compromise security are thus avoided. The aspect of usable security must be considered on the part of consumers when they interact with systems. However, the use of AI systems by users such as financial advisors must also be considered. Here, too, usable security leads to higher efficiency and performance of the systems.

If the user is involved rather than just thinking in terms of technical security requirements, the level of security can be increased, and the motivation, trust and, above all, acceptance of users for the use of AI can be increased in general.

Need 08-13: Procedure for the safety assessment of relevant stakeholders
The majority of issues relating to the management of information security in companies, e.g. ISO/IEC 27001 or the IT-Grundschutz of the BSI, also have an internal scope. The consideration of the need for protection of products and services provided in the application of relevant stakeholders, in particular consumers, is not considered in the aforementioned "information security management system" (ISMS). The approach to be discussed is intended to provide a supportive guideline due to the high individuality of each AI to be used and the associated, ever-changing assessment of criticality, especially in the very sensitive area of financial services.

As early as the idea phase of a new AI the relevant stakeholders have to be identified, their protection needs have to be determined, and appropriate "AI security by design" measures have to be developed.

Example: The client is provided with hardware or software interfaces to AI systems (e.g., software: apps/marketing for investment recommendations; hardware: sensor technology e.g. in the vehicle for telematics tariffs).

Stakeholders of a company are dependent on measures taken by the company with regard to information security and must have confidence in them. Especially in the field of AI, this trust is essential, since AI is usually individual in origin and criticality. Trust can be realized in part through the certification of management requirements such as an ISMS, although products and services that are aimed directly at end consumers, for example, are not covered by this. A standardized approach would open up the view for all stakeholders and, with a transparent approach, create more trust in AI and increase the overall level of security.

In the financial sector in particular, numerous (regulatory) stakeholders can be identified in addition to operators and consumers, and the classification of financial services as critical infrastructures makes trust-building measures particularly relevant.

Need 08-14: Standards for validating the model to assess whether the AI system has been sufficiently validated for use in the production environment
Sufficient generalizability of an AI system must be ensured in order to be able to make reliable decisions in future situations. AI systems are prone to overfitting and underfitting if not adequately developed; therefore, it is highly relevant to validate the model sufficiently to ensure reliable operation in production. Thus, the model must be tested accordingly by adequate methods (including back-testing, stress testing, adversarial attacks) with the goal of a harmonized guideline for testing AI systems. It must be ensured that ML methods that are subject to oversight reviews and approval procedures (internal models for calculating regulatory capital requirements (pillar 1) or in risk management in pillar 2) are suffi-

ciently validated. In order to ensure the appropriate quality, adequate standards must be defined, as existing regulatory requirements currently do not yet take into account the special and complex properties that AI and machine learning technologies will contain in the future.

The particular relevance for the financial sector results from the fact that the models often refer to human behaviour as well as changing environments, e.g., market environments, and cover stress periods. Accordingly, the forecasts must be robust.

Need 08-15: Standards for transparency for error correlation of the system
An AI system should make transparent in a standardized way what the correlation structure of statistical uncertainties is. Statistical uncertainties of the outputs of an AI system are not necessarily independent. A knowledge of the dependency structure is critical for risk management of potential failures of the system. In addition, it must be defined to what extent an input was created under uncertainty (by an upstream model or dataset).

Need 08-16: Definition of sufficient measures of transparency so that the developer knows what additional information needs to be provided in order to construct the appropriate architecture of the AI system
The decision of an AI system must be sufficiently traceable to understand the decision-making process. In addition, transparency requirements should be included in standards, for example by requiring contractors to actively support third-party verification and the creation of traceability. Among other things, the focus should be on understanding what exactly influences the resulting decision, such as being able to explain to a loan applicant, if necessary, the reason for not being granted a loan.

The particular relevance for the financial sector results from the fact that the models often refer to human behaviour as well as changing environments, e.g., market environments, and cover stress periods. Accordingly, the forecasts must be robust.

Need 08-17: Standardization of documentation requirements on the context of origin of models and (training) data
The original context of (training) data as well as finished models must be available at every step of use to ensure verifiability.

If results of individual models are used purely on the basis of their classification, this can lead to unexpected system behaviour. Due to the processes involved in machine learning, marginal data that appears unimportant can lead to undesirable correlation effects. For the use of other model outputs, the original context should be known and taken into account. This includes: Model decisions and weighting, the origin and context of training, validation and test data, the origin and context of real-time input data. Standardization can start here and ensure that no relevant information remains unknown during the acquisition process.

Need 08-18: Standards for transparency on the confidence and model risks of individual decisions
Unlike decisions made with predetermined algorithms, uncertainty about the correctness of the decisions is part of the output of the ML-based AI system. These should therefore be made transparent in a standardized way, e.g. by specifying corresponding probabilities of the possible decisions.

Since several models are often chained in risk management, but these are coupled nonlinearly, knowledge about the error distributions of the individual systems is crucial for estimating the error distribution of the overall system. This is essential for financial service providers as native risk managers.

The Working Group Financial Services ranked the identified needs according to the urgency of their implementation. Figure 46 shows the urgency of implementation, categorized according to the target groups of standardization and research.

Urgency of implementation →

Need	short-term			medium-term			long-term		
Standardization									
08-02	▬								
08-03					▬				
08-04					▬				
08-05	▬								
08-06				▬					
08-07	▬								
08-08					▬				
08-09		▬							
08-10				▬					
08-11							▬		
08-12	▬								
08-13				▬					
08-14				▬					
08-15	▬								
08-16									▬
08-17		▬							
08-18		▬							
Research									
08-01	▬								

Figure 46: Prioritization of needs for the key topic Financial services (Source: Working Group Financial Services)

4.9 Energy and the environment

Artificial intelligence (AI) is advancing into a wide range of application areas. In the integrated field of energy and environment, a complex structure of domain-specific and cross-domain applications can be observed. At the same time, these applications are being used for an increasing range of problems. In the specific aspect of energy systems and technology, the question arises as to what extent AI, as a set of new technologies, can be linked to existing systems and change them. In environmental engineering, AI can support the development of closed-loop processes and decarbonization strategies, as well as provide consumers with feedback on purchasing decisions. Across sectors, environmental aspects are relevant with respect to the further development of energy efficiency as well as the identification of energy needs and environmental impacts of the AI methods themselves. There are AI applications that are explicitly intended to contribute to energy efficiency and environmental protection, such as the optimization of tribological systems (cf. [396]). At the same time, the development and application of AI in any field of application requires energy for the computing power of the technical infrastructure, as well as specific materials and raw materials, which in turn are responsible for environmental impacts in their life cycle.

In order to make the developments and applications of the diverse AI technologies integrative, energy-saving and environmentally friendly and for human benefit, standardization processes are required in many areas. Only a selection of standardization needs can be identified and described here. This chapter focuses on energy technology and environmental impacts. It brings together AI as an innovative technology that is still new to the energy sector with the proven system approaches and application possibilities of the standardization experts in energy technology. The standardization experts have created a functioning architecture in which standards ensure interoperability. The developers of AI, on the other hand, offer ideas and applications that extend this architecture. AI also offers new perspectives for determining environmental impacts by assisting in the management of complex, cross-domain data systems. In the environmental context, AI represents a novel technology set, analogous to the energy domain. Standardization experts and AI developers have designed a common architecture that illustrates the interoperability and need for cross-sector data standards.

4.9.1 Status quo

The supply of energy continues to be a major topic on the political agenda. With the energy transition initiated in Germany and the current dramatic global political changes, a wide variety of goals such as economic efficiency, security of supply, climate protection and the shift towards renewable energies are to be met simultaneously. The smart energy grid, i.e. the combination of energy technology with information and communication technologies (ICT), plays a decisive role here. Standardization, in turn, is a necessary prerequisite for technical implementation and investment security in this area. The advent of AI is now having an immense effect on the status quo in the relevant fields of smart energy grid standardization. This includes the multitude of actors, regional and international activities, and the enormous speed of development. Many of these special features have now been addressed by the activities of the "Smart Energy" system committee (DKE/K901) in the DKE for over ten years. The main effects of AI on established structures will be examined here.

In recent years, a new approach to standardization per se has been established in connection with standardization activities in the field of smart grids, which takes account of the diverse challenges in complex systems. It is essential to integrate a wide variety of sub-areas and the relevant specialist groups. This is achieved by aligning activities with the desired or required services to be offered by the complex smart grid system. Based on these services or functions, a generic model (Smart Grid Architecture Model – SGAM) is used to investigate the implementation options. By describing the services and increasingly detailing them in use cases at the function, information, communication and component levels, the prerequisite is created for the various standards bodies involved to work together on a common goal – the realization of the desired services and functions. This procedure not only ensures coherent standardization work, it also provides the necessary basis for common understanding and consensus building between all parties. In addition, it has succeeded in opening up the collection of basic services and functions far beyond the established circle of participants in standardization.

The impact of AI technologies on environmental engineering presents a complex web. The optimization of systems and processes to maximize energy efficiency and minimize environmental impacts is at the core of AI applications in the environmental domain and a key component in achieving global and national climate change mitigation targets. This concerns, for example, the minimization of friction losses (tribology), path optimization (logistics) and the determination of refurbishment tracks (construction and civil engineering). At the same time, AI represents an energy- and resource-intensive technology set at the meta-level. The energy demand or consumption and the environmental impact of AI applications are thus a fundamental criterion for assessing the quality of artificial intelligence in all areas of application. The joint standardization activities of DIN, DKE and VDI (Association of German Engineers), as well as the joint activities of the European Committee for Standardization (CEN) and the European Committee for Electrotechnical Standardization (CENELEC), have produced a range of environment-related AI use cases. These are flanked in the political environment by position papers, cross-sector publications and regulations from the Federal Ministry for the Environment, Nature Conservation, Nuclear Safety and Consumer Protection (BMUV) and the European Parliament, among others. Here, AI is also seen as a general driver for environmental and sustainability research. The status quo for AI in the environmental sector is presented here in the form of the major standardization efforts, political objectives, association positions and research activities. In view of the complex tension between economic, ecological and social aspects of sustainability, no consideration of socio-technical systems is given here.

Energy technology

Smart energy and smart power grids must be enhanced with real-time data collection, communication, monitoring, and control capabilities to address outages, manage increasingly distributed generation, add renewables and energy storage, while meeting more stringent emissions targets. A largely electrified and automated world requires a continuous, reliable and sustainable supply of electricity. This is achieved through a network that is able to collect and communicate information. Ideally, it is based on standardized hardware, software and processes that ensure seamless integration and interoperability.

Electricity is the ultimate just-in-time product. It is consumed the moment it is generated and must be supplied continuously. During periods of high electricity demand, the plants are extremely stressed. Many of today's power grids were built in the 1960s, sometimes even earlier, and are reaching the end of their useful life. Modernizing networks using the latest technologies is therefore a must. It also helps to improve energy efficiency and to make the generation, transmission and consumption of energy more sustainable. Key technologies used for smart grids include sensors that measure relevant parameters such as temperature, voltage and current; communication systems that enable two-way dialogue with a device; control systems that allow a device to be reconfigured remotely; user interface and decision support systems that provide an overview of asset health and perform advanced data analysis.

Several IEC Technical Committees are developing the standards that will help improve the adaptability of grids to cope with multipath power flows, the integration of renewable energy sources and energy storage, and to become more cost-effective, secure, reliable and flexible. IEC TC 57 develops key standards for smart grid technologies and their integration into existing power grids. Many other IEC TCs contribute to smart grids with standards for sensors, smart switches, automated substations, or smart meters, to name a few. Such standards also serve as a basis for the testing and certification of components, devices and systems. IEC operates four conformity assessment (CA) systems, whose members verify that equipment and systems meet the requirements of IEC standards and specifications. The IEC has established a system committee, SyC Smart Energy, to provide system-level standardization for smart energy and smart grids. SyC helps identify all relevant standards and coordinates the work of the many technical committees involved in smart energy standardization. The IEC has published a roadmap for smart grid standardization that provides guidance for selecting the most appropriate standards and specifications.

The new AI technology must now be integrated into this established system of data models and system architectures. The advantage of existing systems is that the interfaces and processes are already in place and could be used, for example, to make decisions for AI systems.

Environment

Climate protection and decarbonization require comprehensive strategies to reduce energy and resource consumption as well as emissions (environmental impacts). Current approaches for determining these environmental impacts and communicating them to market actors and consumers are characterized by high data requirements and strong data dynamics. Opportunities exist here to maximize energy efficiency, minimize environmental implications, and guide sustainable consumption decisions (see [397]). In contrast, there are inherent environmental risks in the application of AI (cf. [223]). The development and application of Artificial Intelligence and Machine Learning (ML) are in principle characterized by high computational and resource intensity, which puts the added value of environment-related AI and ML applications in tension with their own environmental impacts (cf. [398]). In this respect, decisive and continual monitoring of process-related data is required along the complete life cycle of products and services. This requires uniform data standards and translations (“mappings”) for different data formats and environments.

Research identifies AI as a significant tool in the environmental sector for achieving the 17 Sustainable Development Goals (SDGs) defined by the United Nations (UN) (cf. [399], [400]) and for building sustainable business models (cf. [401]). At the European policy level, AI use potentials are seen in the context of the European Green Deal (cf. [402]). The Policy Department for Economic, Scientific and Quality of Life Policies identifies cross-sector and sector-specific policies. With the help of AI, cross-sectoral behavioural recommendations for market players and consumers to minimize the ecological footprint, as well as measures to reduce the environmental impact of AI itself, are to be developed. In the energy and buildings sector, there is as yet untapped potential for minimizing energy consumption and the associated emissions during the life cycle. In the mobility sector, AI is expected to contribute to the optimization and automation of transportation routes as well as vehicle design. For the agricultural sector, a broadening of focus in AI use from maximizing productivity to reducing fertilizer, water, and land use is recommended (cf. [403]). In European finance, the disclosure requirement of sustainability-related investment indicators “Environmental Social Governance” (ESG) (cf. [404]) in the context of the taxonomy of sustainable activities (cf. [405]) is a challenge to which ML and AI-based systems can make a significant contribution (cf. [406]). The German Federal Ministry BMUV sees AI in the environmental sector as a tool for resource efficiency in industry and SMEs, as well as for processing big data in various sectors of the economy. Furthermore, with the help of AI, the resource-efficient design of AI and ML models will be accompanied and used for the informative scope of energy and environmental indicators of products and services (cf. [407], [398]).

As part of CEN CENELEC’s standardization activities, a Road Map on Artificial Intelligence (AI) has been developed, which contains a range of use cases associated with the Technical Committees (TC).[98] The Roadmap adds to these highly specific issues the need for cross-sector research on AI system architecture, algorithms, and the topic area of ethics, especially the areas of privacy, transparency, accountability, and explainability. Overarching cooperation with other standardization organizations (International Standardization Organization (ISO), IEC) is being discussed (cf. [392]). The joint standardization map on resource efficiency by DIN, DKE and VDI identifies standards and standardization activities for the implementation of the German Resource Efficiency Program (ProgRess III) (cf. [408], [409]). Value creation processes and process chains in all sectors are affected with regard to production and logistics, digitalization, and the communication of energy consumption and environmental impacts. The DIN standards committees NA 172 Principles of Environmental Protection (NAGUS) and NA 005 Sustainable Construction (NABau) have already developed comprehensive specifications on life cycle assessments, associated data documentation formats and product declarations, as well as communication requirements of footprint information (cf. [410], [411], [412], [413], [414]). At the same time, the standards committee NA 043 Information Technology and selected IT Applications (NIA) is working on a range of technical and organizational issues relating to the architecture and use of AI in various committees and joint working committees.

The challenge is to put established environmental and sustainability regulations in a synergistic context with AI-related tools and processes. In this regard, a focus on data environments and the harmonization of data standards from different disciplines can help leverage the potential benefits of AI applications in the environmental domain.

[98] JTC5 – Space, TC61- Safety of household and similar electrical appliances, TC64 – Electrical installations and protection against electric shock, TC134 – Resilient, textile and laminate floor coverings, TC248 – Textiles, TC307 – Blockchain and distributed ledger technologies, TC332 – Laboratory equipment, TC348 – Facility management

4.9.2 Requirements and challenges

The state of the art described in Chapter 4.9.1 shows the current divergence between established systems and technically feasible solutions on the one hand and security architectures and data standards on the other hand. The development of safe and efficient systems in energy and the environment requires an interdisciplinary approach that establishes common standards for safety systems and data formats. In this context, communication from and with stakeholders must be ensured throughout. Due to the wide-ranging nature of the situation, only a selection of specific requirements and challenges can be outlined here. Standardization plays an important role in the planning, construction and operation of new energy and information structures. Cross-industry standardization needs arise in the determination and communication of environmental impacts. Existing standards and specifications from completely different technology areas must be brought together, examined for compatibility and applied in an interdisciplinary manner. Due to new market requirements, new functionalities and interfaces are emerging that will lead to new standards and specifications. This applies not least to the area of interoperability in the field of energy, which must offer experts and laypersons alike access to the optimization tasks as system users. Environmental impacts should be scalable, determinable and communicable in accordance with established methods. An important role for the specification of functions and interfaces to be performed by humans is played by the methodology of use cases. In addition to various description templates for the standardization bodies, structured filing and search functions for use cases are provided. This methodology has become established in the international exchange of information between standards bodies such as: IEC TC 57 and supports the goal of creating a solid basis for the development and expansion of smart energy grids through international standardization. The following selection of use cases (see Table 10) is structured based on DIN EN 62559-2:2016-05; VDE 0175-102:2016-05:2016 [415]. The respective tabular summaries can be found in the Annex (see Table 21 to Table 26 in Annex 13.6).

Table 10: Overview of use cases in the topic area energy/environment

Number	Name	Brief description
1	Autonomous Smart Grid Power Management and Consumption System	Power management system (PMS) and industrial automation and control systems (IACS) are each designed and operated autonomously, but as a coupled system they generate interdependencies that must be controlled and, if necessary, balanced in real time and short time spans.
2	Energy efficiency in buildings and coupling with energy networks	Optimized adaptation of the electricity demand of buildings to forecasted load profiles in generation
3	Personalized AI-powered recommendation systems for sustainable consumption	Personalized, AI-supported recommendation systems for sustainable consumption match product characteristics and individual attitudes and provide precisely fitting product recommendations in various purchasing situations.
4	Scalable determination of environmental impacts in the building sector	Determination of the environmental impacts of buildings and neighbourhoods with adapted level of detail of the data in the Life Cycle Assessment
5	Resource intensity of AI & ML	Integration of a metric/reference method for environmental impacts of AI & ML models in their evaluation.
6	Adversarial resilience learning – Market intervention by aggregators in the distribution grid.	Avoidance of potential attacks on the network in the context of congestion management in volatile load profiles

4.9.2.1 Use case 1: Autonomous Smart Grid Power Management and Consumption System

Autonomous grid systems are widely distributed systems that are animated with mobile data, things or energy flows and equipped with stationary objects, production facilities and buildings. Autonomy in smart grids requires available, stored knowledge about possible critical states or situations to be avoided, and technologies for dynamic control and regulation of components or subgrids. All subsystems in the grid, e.g. grid power management system (PMS) and home, building, industrial automation and control systems IACS, are designed and operated autonomously each by itself. As a coupled system, however, they generate dependencies on each other that must be controlled and, if necessary, balanced in real time and over short time spans. These dependencies influence the stability of the overall system, e.g. high energy demand and low energy input lead to destabilization. In addition, there is resilience to distributed energy resources (DER) component failures that can build up to uncontrollable cascading effects. From the architecture description of the UC SGAM, Smart Grid Reference Architecture model, there are **at least three cooperating systems** [PMS, system interface (SIF), IACS], where each system is represented as a vector of variables. Each assignment of the variable vectors describes a system state, which is changed by incidences. In stable system states of SGAM systems, the variables transport and transform energy. If the system becomes unstable, there is a threat of **blackout incidence**. Observation of critical system state changes, e.g., transition from stable to unstable system state, is one of the tasks of the digital twin, which is equipped with analytical capabilities. The analysis of critical incidences is based on data. All critical incidences must be transparently documented in advance with valid and timely metadata collection. The digital twin uses analytical tools and knows measures for appropriate reactions to blackout incidences that may occur. Blackouts can be used by system control digital twins through data representing ontological system knowledge, and possible misbehaviour can be detected and avoided if necessary. ML technology can be used to search and identify these data for patterns of instability.

4.9.2.2 Use case 2: Energy efficiency in buildings and coupling with energy networks

Since renewable energies are not constantly available, the expansion of renewables requires flexible energy use. With over 40 % of energy consumption, buildings offer great potential for flexible energy use. For example, air conditioning, heating, water heating, or even charging stations for electric vehicles can be used to time electricity consumption in the building and use more energy when it is generated. This can stabilize energy grids and at the same time reduce the carbon footprint of buildings. For this purpose, forecasts must be made for the energy network and its CO_2 factor as well as individual forecasts for energy use in the building.

Based on historical weather and building data, artificial intelligence can determine a forecast for building use. In this way, room occupancy as well as energy consumption can be predicted automatically and, in conjunction with energy network data, an optimized energy utilization plan can be determined. CO_2 and energy savings of up to 40 % are thus possible [416] (see also Figure 47 and Figure 48).

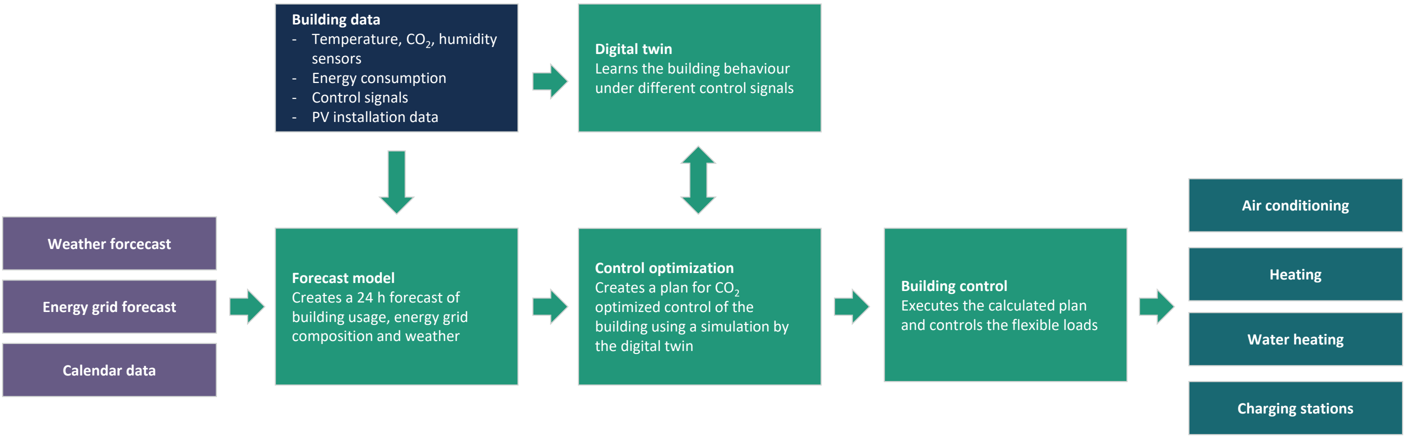

Building data
- Temperature, CO_2, humidity sensors
- Energy consumption
- Control signals
- PV installation data
Digital twin
Learns the building behaviour under different control signals
Weather forcecast
Energy grid forecast
Calendar data
Forecast model
Creates a 24 h forecast of building usage, energy grid composition and weather
Control optimization
Creates a plan for CO_2 optimized control of the building using a simulation by the digital twin
Building control
Executes the calculated plan and controls the flexible loads
Air conditioning
Heating
Water heating
Charging stations

Figure 47: Scheme for optimization and control of buildings (Source: Unetiq GmbH)

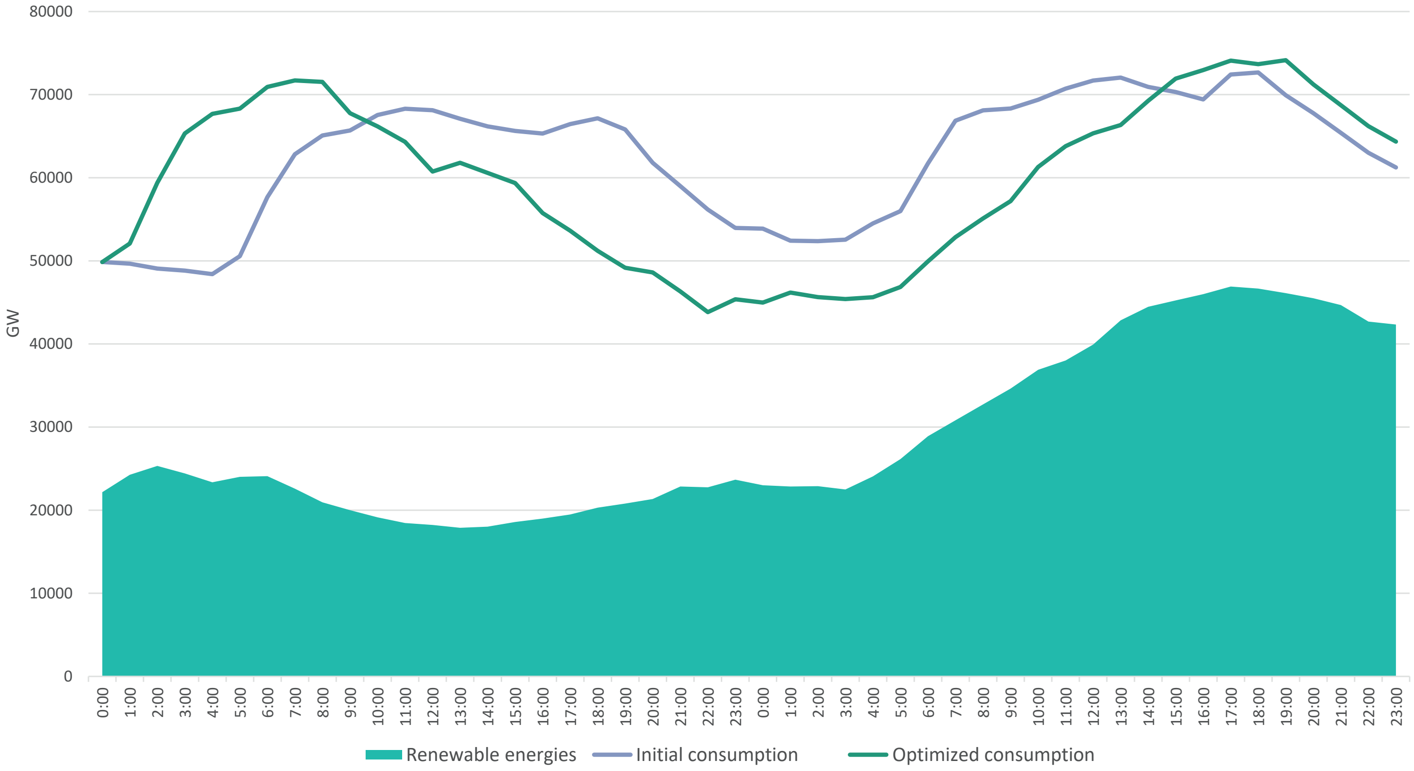


Figure 48: Time history of renewable energy, original and optimized consumption in GWh (Source: Unetiq GmbH)

4.9.2.3 Use case 3: Personalized AI-powered recommendation systems for sustainable consumption

Household consumption and the associated production of goods account for a significant share of global greenhouse gas emissions and global energy and raw material use ([417]).

Consequently, consumption decisions have considerable relevance for the climate, environment and energy. The lack of transparency and clarity of product-related sustainability information when purchasing are barriers to sustainable consumption ([418], [420]).

Personalized, AI-supported recommendation systems for sustainable consumption address this problem by matching product characteristics and individual attitudes and thus making precisely tailored product recommendations. Based on a coherent data basis, which could result e.g., via a Europe-wide harmonized Digital Product Passport (DPP) [421], ([422]), AI enables a more sustainable product selection in various purchasing decisions ([423], [424]). AI could then also capture and analyze personal shopping data over time to provide key insights into shopping behaviour, helping to optimize consumption patterns in line with needs in the medium to long term.

For a broad implementation of such recommendation systems, there is a particular need for standardization aimed at coherence and uniformity of the data basis and AI applications for sustainable consumption. This includes the standardization of environmental indicators used in AI systems and the design of uniform data models and interfaces for the transfer of environmental data between actors along product chains. In addition, efforts should be made to harmonize the algorithms to be used and, in the course of this, to ensure open interfaces and interoperability. Furthermore, with regard to personal data within AI applications, it is important to ensure the quality and protection of data based on data security, ethics and consumer protection standards.

4.9.2.4 Use case 4: Scalable determination of environmental impacts in the building sector

Climate protection targets require the development of sector-specific decarbonization strategies. In the building sector, physical and technical refurbishment tracks must be defined for existing buildings, and climate-neutral standards must be set for new buildings. This requires a comprehensive sustainability assessment and estimation of the decarbonization potential of components and materials. Life Cycle Analyses or Life Cycle Assessments are used to determine such energy requirements and environmental inputs. There are a broad practice and established standards for Life Cycle Assessment, which are basically characterized by a high demand for information and time (cf. [425]). At the same time, the long process chain in the complete life cycle with many influences and variables implies a high volatility of the assessment results.

Based on building physics and systems engineering preferences, AI applications can in principle formulate proposed solutions for the climate-neutral planning of buildings and neighbourhoods. This requires a continual learning system in the background that uses the processing of relevant building and neighbourhood data to determine the environmental inputs in the life cycle.

4.9.2.5 Cross-section use case 5: Resource intensity of AI & ML

Artificial Intelligence and Machine Learning are used to find solutions and increase efficiency in a wide range of areas. In principle, AI and ML models are characterized by high computational runtime and performance, which in turn imply high energy consumption and environmental impacts (cf. [398]). The additional benefits of AI and ML applications thus conflict with their resource consumption. This applies in particular to such applications that are intended to increase resource efficiency and reduce environmental inputs. From a technical perspective, resource consumption depends in principle on the data requirements and runtime of the algorithm.

At the higher level, AI can provide feedback on the sustainability assessment of AI and ML applications. This requires a metric or reference method for measuring and comparing the performance of algorithms.

4.9.2.6 Use case 6: Adversarial resilience learning – Market intervention by aggregators in the distribution grid

In distribution grids, one future challenge posed primarily by the energy transition is congestion management. The change in load flow means that even at the lowest level, “prosumers” no longer just consume (electric) energy, but also actively feed it into the grid. The previous expansion of the grids as well as the operational planning did not have this aspect as an original focus. In the probable case of a higher feed-in than consumption, reverse flows occur which can cause bottlenecks – conversely, voltage problems can also occur due to numerous new consumers. In short, the instrument of congestion management becomes relevant. These bottlenecks can indeed occur randomly, but can also be targeted through accords and must be addressed. An AI can detect attacks and impending bottlenecks, identify gamification, and be used as both an attack detection and grid planning tool.

4.9.3 Standardization needs

The use cases in Chapter 4.9.2 imply concrete needs in the standardization of processes and formats. There is also a need for fundamental research involving various disciplines in industry and science. The following remarks describe standardization needs, the fulfilment of which will flank the use cases presented and contribute significantly to their success. Due to the selection of particularly urgent needs and challenges already made, the following aspects can also only represent an excerpt from the complete range of needs. The ability to connect to further requirements is therefore dependent on the progressive development in politics, research and standardization.

Need 09-01: Interoperability of terminology, semantics, taxonomy and data
Material science and economics are confronted with fundamental issues to increase resource and energy efficiency. This applies in particular to the field of tribology, since friction and wear optimization have a direct impact on material and energy consumption. Inconsistencies in terminology and dependencies arise in characterization and modelling methods due to many domains involved The **FAIR** principles (Findable, Accessible, Interoperable, Reusable) must form the basis for action here. Creating or harmonizing terminologies, semantics, and taxonomies across domains can ultimately only be done through stakeholder engagement and requires consen-

sus-based exchange. Furthermore, to evaluate the reliability of AI decisions, one must consider integrating appropriate metadata (e.g., sensor type and measurement imprecision for sensor data) into the data models. Consequently, this process should be subject to regular review and accompanied by standardization.

Need 09-02: Schemas and mapping for GIS/BIM integration
To determine the environmental impacts or **L**ife **C**ycle **A**ssessment (**LCA**) in construction, a high demand for data arises at the building and especially at the neighbourhood level, which must be served efficiently. **G**eographic **I**nformation **S**ystems (**GIS**) and **B**uilding **I**nformation **M**odelling (**BIM**) overlap as common modelling methods. In particular, GIS-based building models at **L**evel of **D**etails (**LoD**) 3 and 4 have qualitatively similar information to detailed, BIM-based building models. Using data from both domains can provide significant leverage in environment-related artificial intelligence and machine learning applications. However, this requires a common data standard in the form of model translations, mappings of data formats and database schemas. Such a designed data standard should be continuously accompanied and receive regular updates as a result of updates from both domains (especially **OGC** (**O**pen **G**eospatial **C**onsortium) for GIS and buildingSMART for BIM).

Need 09-03: Coherence and uniformity of databases and AI applications for sustainable consumption
The uniform, cross-sectoral or sector-independent disclosure of environmental impacts and circularity of goods and services requires a common format for communication. This includes a common data standard for broadly determining environmental impacts. This standard and integrative data formats simplify the construction of AI-based recommendation systems for sustainable consumption. Specifically, there is a need for the standardization of product databases, associated database schemas, and data mappings to ensure interoperability. Furthermore, a learning feedback system or the continuous optimization of algorithms requires a formulation of the possible uses of data on personal consumption behaviour that complies with data protection requirements. The aspects outlined involve a range of stakeholders from industry and science who should be involved in the standardization processes.

Need 09-04: Methodology for determining the environmental impact and performance of artificial intelligence and machine learning models
The fundamentally high data and computational intensity of AI and ML models imply high energy consumption and environmental impacts, which are in principle in tension with the benefits of such models. Furthermore, the performance of such models is highly dependent on the use case. Depending on the use case, different algorithms have different runtimes and accuracies. A systematic capture of these characteristics as meta-parameters enables a priori selection of appropriate algorithms for AI and ML applications within the categories of supervised/unsupervised/reinforcement learning and should be accompanied by standardization. To determine the sustainability of such systems, a uniform standard with measurable evaluation criteria is needed. To date, no standardized procedures exist for this purpose. It must first be determined whether an absolute metric with specific measurement criteria or a reference procedure with a standardized reference system will lead to better evaluability and comparability. The selected procedure is then to be formulated to the extent that an AI-supported feedback system can be built to assess the runtime, accuracy, and sustainability of AI and ML approaches.

Need 09-05: Input formats for learning systems
In the context of domain-specific processes, it is principally noticeable that knowledge has to be laboriously prepared and reformatted for AI. Formats must be established as a standard in order to make a broad base of knowledge available in such a way that it can be used in numerous applications, thus establishing “growing” knowledge. A unified semantics as well as syntax, similar to the agreement on a business language, enable quick access to the documented knowledge as well as better reuse.

Need 09-06: Overview and reference model building
Standardization and coordination of content in standardization bodies regarding definitions and taxonomies leads to a common domain semantics. To this end, leading bodies are to be defined for specific topics and content. The creation of a standards map provides easy and quick access to the complex interdependencies of AI in the context of each of the leading bodies, and can therefore support the use of standards by making contact persons and knowledge carriers more accessible.

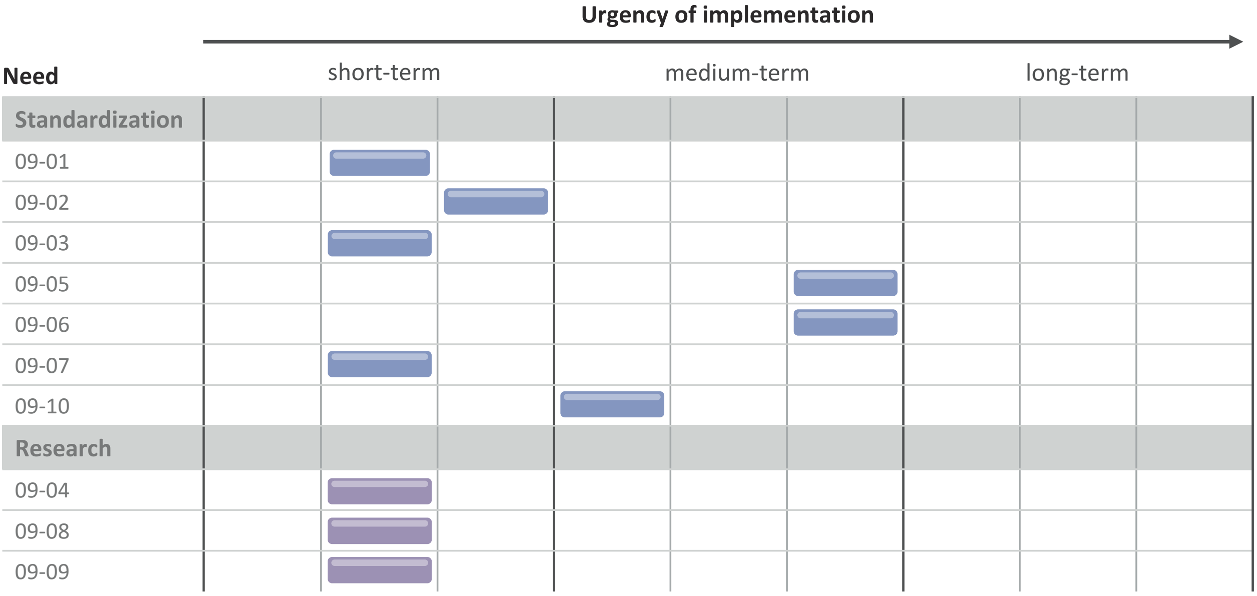
Need 09-07: Dimensioning and conceptualization of I4.0 reference architecture models (RAMs)
The RAMs for smart manufacturing (SM), smart grid (SG), and other technical infrastructures are usually cubic models that use comparable categories such as communication layers, value stream states, and usage or production hierarchies. However, the RAM terms and concepts are not all aligned due to the disjointed application domains of SM and SG. This results in a need to compare and align the RAM terminologies used from the application domains semantically, functionally, security-wise, and ethically.

Need 09-08: Dynamization of the static reference architecture models (RAM)
Smart manufacturing (SM), the smart grid (SG), etc. are architecture models. Therefore, no means are provided to model dynamic processes as part of the static structures. Consequently, there is no conceptual definition of a process variable in today's RAM. However, a system-of-systems (SoS) is a communicating multi-variable system that requires the resources of a RAM in the value stream for the transfer functions it contains. The need for action that can be derived from this is the integration of all means for representing architectural (static) structures and variable (dynamic) behaviour in dynamized RAMs (dRAMs).

Need 09-09: Digital twin to control and verification tasks in Smart Grid Architecture Model (SGAM) systems
In SGAM systems, the risk of DER equipment failures (outages) is particularly high after natural events, such as thunderstorms and storms, because in most cases a cascading load shift occurs that leads to an overload in the undetected weak points and consequently causes outages of planar power supplies. In this context, a digital twin will collect operational and load data on DER in an SGAM or RAM-I4.0 system in order to transfer them into an operating system model. Here, dangerous incidences are to be analyzed. To proactively avoid such incidences, the digital twin could simulate load shifts to avoid overload and weak points simultaneously with weather patterns.

Need 09-10: Calculation method for determining the CO_2 factor from the electricity mix
To determine CO_2 emissions from electricity consumption at a given point in time, it is necessary to allocate emissions to the kWh generated. Current standardized procedures for allocating these emissions provide for a static calculation based on a fixed factor, which is updated as necessary with new editions of the standard. This methodology does not sufficiently account for the volatility of the electricity mix, as weather-related fluctuations in generation from renewable sources cannot be taken into account. Thus, an agile calculation method with ahigher temporal and possibly geographic resolution is needed to more accurately determine the environmental impacts of electricity consumption.

The Working Group Energy and the Environment has ranked the identified needs according to the urgency of their implementation. Figure 49 shows the urgency of implementation, categorized according to the target groups of standardization and research.



Need	short-term			medium-term			long-term		
Standardization									
09-01		■							
09-02			■						
09-03		■							
09-05						■			
09-06						■			
09-07		■							
09-10				■					
Research									
09-04		■							
09-08		■							
09-09		■							

Figure 49: Prioritization of needs for the key topic Energy and the Environment (Source: Working Group Energy and the Environment)

5 Requirements for the development and use of standards and specifications

The rapidly advancing technology development and industrial application of AI systems currently poses a major challenge for the development and use of standards and specifications. Different industries use different AI technologies that are specific to the use case, depending on the field of application of the AI solution. In most cases, the specifics of the application are met by state-of-the-art approaches from AI sub-disciplines, which are continually adapted and refined. Consequently, the dynamics at the interface between AI research and industrial development and application are particularly high.

Standardization must take account of the dynamics between applied research and mature industry development and pursue new approaches to analyzing standardization needs, developing market-ready standards and specifications, and reviewing and adapting existing standards and specifications.

Various approaches are currently being taken to address these challenges, as outlined below.

5.1 AI readiness of standards

Artificial intelligence (AI) is penetrating more and more areas of everyday life and industrial applications. At the same time, it must be assumed that AI can only develop its full potential if it is used in accordance with recognized quality benchmarks, which ensure that AI applications are safe and reliable and that they are used in accordance with societal values. Standards and specifications are a tried and tested way of implementing such quality benchmarks. An “AI made in Europe” brand based on high-quality standards and specifications can be a key competitive factor for the German and European economies. This requires a review and, if necessary, adaptation of all existing standards and specifications with regard to AI, as explicitly called for by the German government’s AI Strategy[99] [2] in Field of action 10.

To this end, a project to evaluate the “AI readiness of standards”[100] was launched in January 2022 under the leadership of DIN and on behalf of the German Federal Ministry for Economic Affairs and Climate Action (BMWK). With a project duration of two years, the project will initially run until December 2023. The Fraunhofer Alliance “Big Data and Artificial Intelligence”, Beuth Verlag and DIN Software are also involved.

The project has two focuses and can be divided into the following aspects (see Figure 50):

1. the relation of the content of the standards to AI technologies
2. the machine executability/readability of the standards themselves

The machine executability of standards is already being investigated by the SMART Standards initiative (see Chapter 5.3). The “AI Readiness of Standards” project is intended to build the bridge to AI-specific use cases and derive any requirements for machine interpretability or future potential uses. The knowledge gained from the project will additionally be incorporated into the “Initiative Digital Standards” (IDiS) network and create synergies.

The other and far greater focus within the project (aspect 1) is on the content of the standards. The starting point is the assumption that artificial intelligence will sooner or later be of great significance for all economic and social areas. Standards and specifications exist for almost all industrial sectors and fields of application. Currently, the German body of standards comprises more than 30,000 standards (DIN, DIN EN, DIN EN ISO/IEC). This means that a large proportion of existing standards are likely to have points of contact with AI technologies and therefore need to be reviewed accordingly and supplemented with AI aspects. At the same time, as of today there is no central overview of which standards are designed for the use of AI technologies in their field of application.

The project is intended to be an inventory of the cross-section of the entire body of standards to answer questions such as:

→ How many and which standards have points of contact with AI technologies?

→ Which of these standards are already prepared for the use of AI?

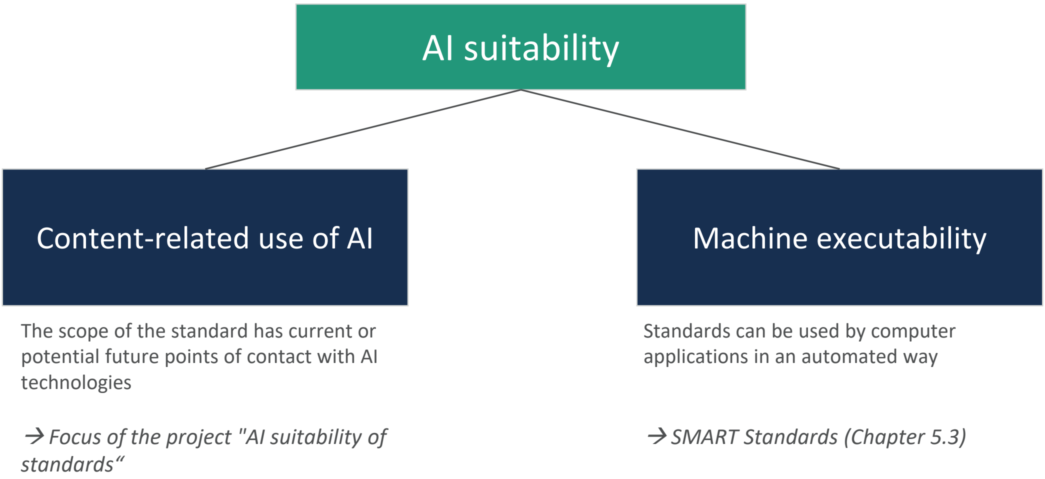
→ Which standards need to be revised in a timely manner in this regard and how could this revision be carried out?

The goal is to develop a scalable methodology and a prototypical AI tool (software) that can be applied to the entire set of standards in perspective.

[99] https://www.ki-strategie-deutschland.de/home.html

[100] https://www.din.de/de/din-und-seine-partner/presse/mitteilungen/din-startet-projekt-ki-tauglichkeit-von-normen--872810

Figure 50: Structure of the project "AI readiness of standards" the Environment (Source: DIN)


AI suitability
Content-related use of AI
The scope of the standard has current or potential future points of contact with AI technologies
→ Focus of the project "AI suitability of standards"
Machine executability
Standards can be used by computer applications in an automated way
→ SMART Standards (Chapter 5.3)

Project procedure
In the context of the project, a definition of AI readiness (in terms of content) will first be developed. The concept of AI readiness will be refined and made operationalizable, for example, by establishing evaluation criteria for determining AI readiness. To elaborate the definition and methodology, two aspects are particularly important: the technical or standardization knowledge and the expertise on AI methods. The two groups of experts from the standardization bodies on the one hand and the AI experts from the Fraunhofer-Gesellschaft on the other hand are to integrate both points of view. The "AI readiness of standards" project has already been presented at the advisory board meetings of some standards committees. All interested specialists are invited to participate at any point in the definition, methodology or pilot project workshops.

In analyzing the examples, special attention is to be paid to the case where standards and specifications restrict the use of AI. A distinction must be made here between fundamental inadequacies of existing AI methods, which prevent their use according to the current state of the art, and the case where an expansion or further development of standards in terms of content can overcome the restriction. A possible example here is the topic of "safety", the requirements of which currently prevent the use of AI for certain fields of application.

The first key outcome is a meaningful definition of "AI readiness of standards," in terms of content. In order for this to be applicable in the various areas of standardization, an operationalization follows in the form of a working aid, which allows the corresponding experts to make an independent assessment of whether a standard is suitable for AI.

The review of standards for AI readiness is to be additionally facilitated by a machine-supported process. To this end, a prototype AI tool will be developed to assist in the selection and evaluation of standards. The level of support here depends on the number of reviewed standards with which the AI tool is trained. The more standards that are manually reviewed, the better the AI tool will be able to evaluate.

As part of pilot projects and as a basis for the development of such an AI tool, the next step will be to assess the body of standards with regard to selected specialist areas with a particular connection to AI (for example sectors such as mechanical engineering, automotive engineering, medicine or cross-cutting topics such as ethics and safety, analogous to the key topics of the German Standardization Roadmap AI). In further steps, various industry and AI experts will analyze the standards for these selected specialist areas with regard to their AI readiness on the basis of the developed working aid.

The project results will be:

- → a methodology for assessing AI readiness (possibly extendable to other areas of application, „climate protection readiness", etc.);
- → a list of reviewed and evaluated standards;
- → recommendations for revisions to the standards that have been assessed as not being suitable for AI;
- → the prototype AI tool developed for machine-assisted evaluation of the remaining body of standards.

Through the project, the participating specialists will be sensitized to the topic of AI and, with the support of the AI experts, will be able to identify and evaluate points of contact to their priorities and their work programme and to set impulses. The AI experts will be available for the duration of the project and can be called upon for advice. Through the AI experts and their network, new AI experts can also be recruited for the committees, thus making a significant contribution to the consideration of new technological developments in standardization. The AI tool will also provide an additional valuable technical capability to support the standards committees beyond the life of the project. The project can thus make a very significant contribution to strengthening Germany as an industrial nation.

5.2 Agile development of standards and specifications

The high level of dynamism in the development of AI technology has an impact on the requirements placed on the process of developing standards. Agile approaches and processes are needed here, which constantly incorporate reciprocal impulses from experts in the design of standards and specifications and support the collaborative development of standards and specifications.

At the heart of this approach is the XML file format, which has established itself as an integral part of the further processing of standards and standards content and is a key pillar of the efforts surrounding SMART standards (see Chapter 5.3). The XML-first strategy of the DIN Group provides for the earliest possible integration of the XML file format in the standards development process. According to the strategy, standards and specifications are to be created directly in XML in the future, thus enabling, among other things, the successive replacement of conventional text processing software, downstream conversion processes, and existing media discontinuities. On the basis of this conversion, standards development processes can be designed more efficiently, and potential sources of error, which sometimes generate considerable additional costs, can be reduced.

A suitable tool is needed to realize this project. With FontoXML, the focus is on an XML editor that enables both direct creation in XML and collaborative editing of content, both of which are a novelty for standards development at DIN and DKE.

The XML editor will provide standards writers with a technical basis for future content capture and further advance the digitization of standards production. At the same time, the collaborative development of content will further strengthen the transparency of the standardization process.

The experts in the standards committees will be able to work on content in parallel. This will significantly increase the agility of the processes for developing and revising standards and specifications. Comments on work statuses can be submitted via the tool, directly viewed and assessed by others. Current work statuses are always stored in the cloud where they can be revised, so that standardization bodies no longer need to manage offline copies. Content is therefore not lost and the tedious management of multiple Word documents is also eliminated.

Permanent access to the current status of work allows, among other things, a faster response to changing framework conditions in the course of a standardization project and enables greater flexibility in processing for both the experts and the standardization body. Ultimately, these advantages can also contribute to shorter development or revision times and thus to faster availability of standards and specifications.

The international (ISO and IEC) and European standards organizations (CEN and CENELEC) will also rely on XML in the future and plan to enable the development of standards content using FontoXML directly in XML format. Since the end of 2020, the XML editor has been tested in pilot projects and further developed together with the developers.

5.3 SMART Standards

New design of standards and specifications for integration into AI application processes
The direct further utilization of standards and their contents in downstream processes is gaining increasing attention. Companies will expect efficiency gains[101] in the future from standard components (value tables, part descriptions, 3-D models, software, requirements definitions, test methods) that can be adopted and executed directly by machines.

To achieve this goal, DIN and DKE have been working on the SMART Standards project for several years.

[101] See C. Wischhöfer, P. Rauh, Standards of the Future – Stand der Arbeiten zum Thema maschinenausführbarer Normeninhalte. DIN-Mitteilungen, August 2019, pp. 4–8.

A SMART standard is understood to be a **s**tandard whose contents are **a**pplicable and **r**eadable for **m**achines, software or other automated systems and can also be made available digitally in an application/user-specific manner (**t**ransferable).

The following shows the developments that have taken place since 2020. For this purpose, the initial situation in 2020 and the current status (2022) are presented.

Initial situation 2020

The workflow, which has been established for decades, functions successfully and in a balanced manner on the basis of agreements between the process partners involved. The underlying principles are carefully coordinated in compliance with standards and legal requirements and guarantee reliable management of standardization results in customer-oriented systems.

The upcoming profound procedural changes in the context of SMART Standards elaboration, content management, distribution and use will have to be delineated and redefined against the background of existing established and regulated practices. The decisive value ("asset") of a standardization subject must be preserved.[102]

Current status 2022 and further work

The essential requirement of a workflow for SMART standards is to develop and provide structured and semantically enriched content that is the basis for machine processing, in particular also for AI applications.

The 1st edition of the Standardization Roadmap AI described the essential development steps of a future standardization process (content creation) and identified a profound need for change in the process.[103] In addition to content creation, the content management and content delivery process steps must also be further developed so that they can process fragmented and semantically enriched standard content and deliver it to the application (content usage).[104]

The projects on SMART standards at European and international level have dealt intensively with the challenges outlined above over the past two years. They have each set up their own project structures for this purpose, either working together directly or exchanging and coordinating information through formal and informal channels:

At CEN-CENELEC, **Workstream 3** "Technical Solution" develops the information model and the technological infrastructure for SMART standards. Content capture tools (such as XML editors for the creation of Level 3 content based on FONTO technology) play a central role here, as they are a mandatory technological prerequisite for achieving a higher degree of structuring while at the same time maximizing process efficiency.

Workstream 4 "Operationalization" describes the content creation process down to the level of new workflows with a future capture tool and defines the requirements for the supporting organization. These three central components – process, organization, and technology – are currently being prototyped so that they can be piloted and tested in specific standards projects (in coordination with **Workstream 1** "Standards User Engagement") in 2023.

In **Workstream 5** "Business Model", new forms of delivery (in the sense of content delivery) are being evaluated according to the recorded use cases (**Workstream 2** "Standards Maker Engagement") and derivations are being made with regard to commercial and legal aspects (e.g. licensing and terms of use).

At the international level, ISO has formed analogous working groups with the "subgroups" within ISO SMART, and IEC with the "task forces" of SG 12. These working groups are characterized by a large overlap with the European project, both in terms of content and project participants. This ensures a transfer of know-how from the European to the international level (and vice versa), which is crucial if a model for Level 3 is to be made available for productive use by 2024, and if joint further development towards Level 4 content is subsequently to take place.

[102] See Standardization Roadmap AI (1st edition), Chapter 5.2.2 [63].

[103] See Standardization Roadmap AI (1st edition), Chapter 11.4.3, Annex "Top-down method" [63].

[104] See Standardization Roadmap AI (1st edition), Figure 31 and Figure 37 [63].

Initial situation 2020
One challenge will be to consolidate a common understanding among developers and users of SMART standards.

The stage model must be verifiable and adaptable to other models, e.g. Reference Architecture Model Industrie 4.0 (RAMI4.0) /HHHD-17/.[105]

Current status 2022 and further work
In the white paper "Scenarios for digitizing standardization and standards" from IDiS (Initiative Digital Standards), the IEC Utility Model (also the SMART Standards Level Model) was expanded to include a Level 5: Machine-controllable content. The content of a standard can be amended by machines working unassisted, and adopted by automated (distributed) decision-making processes The content adopted in this way is automatically reviewed and published via the publication channels of the standardization organizations.

The white paper goes on to explain that AI applications benefit from improved machine applicability because this increases the interpretability and evaluability of normative content.

Thus, the white paper supports the thesis that SMART standards introduce rules and processes in the description of content that make it easier for AI applications to better process the content thus captured. Level 5, mentioned above, goes one step further and describes the possibility that AI applications themselves can become part of decision-making processes (e.g., the standardization process) and thus act as active participants.

In IDiS a pilot project was started at the end of 2021 (duration approx. twelve months), which aims to develop a domain-specific language model based on approx. 40,000 DIN and VDE standards. This language model is to be used to identify suitable text passages from relevant standards for a product whose characteristics are defined in the ECLASS product data standard. In this context, the pilot project investigates the general suitability of normative content for the applicability of machine learning methods and the extent to which AI methods can support the search for relevant data.

Initial situation 2020
In all committees, an important new aspect is repeatedly addressed: How do we prepare the actors along the entire value creation process for the new requirements?

Another aspect of systemic relevance for the future concerns the definition of the requirements for the changed qualifications not only of the external, but also the DIN-internal "actors" in the overall process. The existing concepts form the basis for defining the requirements for jobs as well as further training opportunities and must consequently be further developed to describe the new tasks arising in SMART Standards processes for all those involved in the process.[106]

Current status 2022 and further work
The process-related, technological and business model-related SMART standard aspects are being intensively discussed nationally at DIN/DKE, European-wide at CEN/CENELEC and internationally at ISO/IEC in various working groups, and initial solutions are being developed. At present, the changed requirements for process participants along the entire value creation process

Content Creation > Content Management > Content Delivery > Content Usage

are, however, not being adequately taken into account. Planning and training alone will not ensure change by the stakeholders. With regard to the future use of AI-supported processes, it is not sufficient to offer only a few experts a platform for their considerations and realizations.

The future requirements for the complete execution of the various sub-processes and associated tasks are different – and so are the requirements for the people (or the availability of the necessary competencies in corresponding degrees of proficiency) who have to process the tasks. The methodology is known and is currently being further elaborated and realized in DIN: "Functional descriptions" with clearly defined scopes of action and value-forming task descriptions are a prerequisite for comprehensibly describing and specifying the future requirements in content development, presentation and use in companies, identifying and strategically anchoring the necessary positions in the company, and training people on the job in the sense of building up competencies.

[105] See Standardization Roadmap AI (1st edition), Chapter 5.2.3 [63].

[106] See Standardization Roadmap AI (1st edition), Chapter 5.2.3 [63].

In industry, initial functional descriptions already exist along the value chain considered here. Experience has shown that the attractiveness of these positions for external applicants is very high.

Initial situation 2020
SMART standards are one of many knowledge domains and basically enable AI systems to automatically and optimally use the information they contain in the various sub-processes in a company.

The design of the necessary data models and interfaces will have to be part of this project and thus makes an important contribution to the further penetration of AI applications in the sub-processes of companies.[107]

Current status 2022 and further work
The IEC Utility Model with its five levels (Level 0 – 4) was further discussed in ISO and IEC and accepted as a common basis for describing the basic machine applicability of SMART standards.

In addition, further concepts for the future use of SMART standards were developed on this basis, which are currently being further discussed and developed in the ISO and IEC working groups (ISO SMART, IEC SG 12).

For example, there are initial ideas of an SAM (standard architecture model) and an SAS (standard administration shell). Both concepts are based on Industrie 4.0 ideas (RAMI 4.0 (Reference Architecture Model Industrie 4.0) and AAS) and should help to better classify and discuss the functionalities and responsibilities around SMART standards. Following the RAMI model, the SAM assigns activities and functions of SMART standards to different dimensions (application layer, utility level and standard life cycle) in order to further improve understanding and differentiation between applications. The SAS, on the other hand, is more of a technical model and describes how functions and responsibilities can be divided to provide consistent access to SMART Standards content. In IDiS (Initiative Digital Standards), the national community for SMART Standards, a first pilot project (duration approx. 15 months) on the topic of the administration shell and submodel of a digital standard started in mid-2021.

Initial situation 2020
The economic benefits of standardization have been quantified in some countries. In Germany standardization saves the economy 17 billion euros per year. The quantification of an economic benefit of SMART standards is not yet available and can so far only be mentioned qualitatively.

Within the framework of the project, an economic evaluation must be carried out with regard to effort, benefits, realization period, quality, etc. of the various approaches. Afterwards or during the project, a prioritization of the approaches can be made.[108]

Current status 2022 and further work
The contribution to German economic growth of the current stock of standards is around 17 billion euros per year, which is roughly equivalent to 0,7% of gross domestic product.

The more standard content that can be tapped automatically through SMART standards, the higher the current share is likely to be, especially for the use phase in the value creation process. It is obvious that the potential for increasing efficiency in the application of standards through such automated and application-specific provision and transfer of standards information is considerable.

Since the standards organizations are currently unable to quantify this potential, DIN and DKE are planning to investigate this within the Initiative Digital Standards (IDiS) as part of a project.

Initial situation 2020
The focus on IT-based processes and their further development in “content management” and “content delivery” offers the opportunity to quickly arrive at concrete solutions that provide valuable input for Level 4 (AI).

For downstream AI application processes, this means: A validation of the accuracy of the automatically determined (partial) information [today: fragmented standard content] must be performed. Learned empirical knowledge can provide essential support for the evaluation.

[107] See Standardization Roadmap AI (1st edition), Chapter 5.2.4 [63].

[108] See Standardization Roadmap AI (1st edition), Chapter 5.2.5 [63].

General rules for describing the fragmented contents of standards in standards, as well as the methodical elaboration of the exact locations of use (impact locations), do not yet exist for this approach and must be elaborated. In order to provide AI application processes with fragmented standard content in a scalable manner, appropriate provisions must be agreed upon.[109]

Current status 2022 and further work

Users of standards documents often invest a lot of time in research in order to extract relevant information from standards (e.g. requirements, formulae, product and classification features) and to be able to use it. The large number of potentially relevant standards makes the research effort more difficult. Systems such as the Semantic Standards Information Framework (SNIF) provide good support in this regard and facilitate keyword and topic searches. However, applications like SNIF are based on fixed rules and keywords. They are limited to these and thus lead to predefined results.

Modern methods from the field of artificial intelligence, specifically in this case natural language processing (NLP), are the basis for strong improvements in the language understanding of machines in various domains. For this purpose, pre-trained language models (e.g. German BERT) are used, which are trained on a wide variety of texts. Pre-training the models gives them a basic understanding of the domain from which the texts and the information they contain originate.

Language models based on international and German-language standards are trained in various projects. These pre-trained language models can be refined for different use cases. One of these use cases is, for example, the extraction of relevant standard content (e.g. requirements or product features). Furthermore, datasets are created that contain questions to standards in combination with relevant text passages of the standards as answers. Thus, pre-trained language models can be refined in the form of a specialized model such that they learn to identify and extract appropriate text passages to a question. Moreover, statistical classifiers can identify relevant standards content based on rule-based approaches. Thus, among other things, text passages that do not represent requirements in terms of content, for example, can probably be rejected.

Initial situation 2020

Based on XML-converted documents and in compliance with the NISO STS, the “con:text” service was developed, which can be linked to various standards management systems. The function set aims to capture the content in greater depth, to play out relationships simultaneously and to make them visible to the user in a user-friendly way via numerous functions.

The function set of con:text reflects the requirements of the users. Thus, application know-how is created here that can be relevant for the function formation of AI application processes. At the same time, the con:text application can benefit from the results of the AI project. Collaboration on the AI project should be made possible.[110]

Current status 2022 and further work

The goal of the previous project was to develop an online editor that could be functionally integrated as seamlessly as possible into the existing split-screen interface of con:text.

Content is now captured here in an HTML interface and structured directly in XML, so that the underlying schemas (for example, NISO-STS) make subsequent complex conversion processes superfluous. Content creation can be done in project teams with different responsibilities (initiate, edit, release, etc.). The online editor can flexibly map these roles or adopt existing rights and role concepts from third-party systems via an interface connection.

Future AI-based measures for user support include the (partially) automated checking of text- and data-based content to compare requirements, values, value ranges, or other provisions; searching for and finding process-relevant text passages in the standard currently being processed, in cited standards, or in standards that match the topic; selecting and extracting defined components such as mathematical formulae, tables, or requirements; and transferring such search results into structured output formats (including ReqIF). As part of AI-based further development, the aforementioned functions can also be provided in products supported with con:text, such as standards management solutions, online services or portal services.

[109] See Standardization Roadmap AI (1st edition), Chapter 11.4.1 [63].

[110] See Standardization Roadmap AI (1st edition), Chapter 11.4.1 [63].

Initial situation 2020

The solution is to automatically extract standards content and convert it into a machine-executable knowledge representation form that can be accessed by different authoring systems. From the knowledge that can be gained during the concrete concept implementation, requirements and design rules can be derived to a higher abstraction level of the "next generation standard".

The restructuring of the existing, very large body of standards comes up against capacity limits and would only be economically justifiable for defined subject areas. Here, an application of artificial intelligence in the extraction phase of the bottom-up approach is to be investigated in order to support this work step by machine.[111]

Current status 2022 and further work

In addition to full-text standards, the provision of customer-specific and requirements-specific partial content is also increasingly in focus. In this context, different elements of standards present different challenges that need to be addressed and worked out individually. The decisive factor here is the preparation in data structures that are as generic and broadly applicable as possible (i.e., approaches for storing, arranging, and linking data that are as transferable as possible) as well as a broadly based metadata strategy (i.e., definition of a comprehensive and target-oriented description of the relevant data in order to make it optimally identifiable and selectable for different use cases) – this forms the basis for further usage scenarios.

Formula content in particular, where labelling, individual transferability and cross-document networking are important, lends itself to this. Since formulae are already marked as XML elements, they can be quickly identified. This is a good basis for further processing.

Formulae can also be provided with additional information beyond their own representation. When mathematical, physical or chemical relationships are represented by formulae, additional information is often presented outside the formula itself in the context of a standard. These can be additional extending or restricting properties, which are placed in the surrounding text accompanying the formula. Likewise, alternative or supplementary descriptions of the same or related subject matter noted in the same or other standards are possible.

This information can expand on or give detail to the meaning of a formula depending on the context. Their correct and case-related evaluation and observance thus represents a major challenge. It also follows from this that, although formulae in standards have great relevance in themselves for understanding the relationships expressed by them, in their actual representation in standards they cannot be detached from the context surrounding them.

The need of users to simplify the handling of formulae in standards in everyday practice is made clear by the large number of software offerings that have specialized in practice-related support for the use of formulae for a wide variety of applications.

If it is possible to work out new, preferably automated extraction and semantic processes on the basis of formulae, these can be carried over to other types of content. Important here are stringent content publishing policies, as flexible as possible data structures based initially on XML, or comparable environments that enable the creation of scalable content databases.

The comprehensive consideration of formulae, their optimal storage, context-related linking and the process- and system-oriented presentation forms an important approach for the modelling and provision of all further standard contents.

AI-based methods represent the most promising approach to creating the required data structures, as this is the only way to provide the desired SMART standard content in a timely manner.

Initial situation 2020

The largely automated and AI-supported overall process requires integrated, overarching action on the part of those responsible for the process, so that previous boundaries of responsibility must be reconsidered and redefined. Most definitely, the content responsibility for "content creation" must be located in the process of developing the standards – the primary content. Postprocessing in the sense of a subsequent interpretation of content for further processing must no longer exist.

Standardization phase: Currently, the language (prose) of subject matter experts cannot be directly transformed into

[111] See Standardization Roadmap AI (1st edition), Chapter 11.4.2 [63].

a machine-interpretable form in terms of SMART standards. With future available experience and learned knowledge in AI application processes, it is nevertheless to be conceived that an AI-oriented modelling can be realized.

Formalization and IV. modelling: Transformation using "semantic triplets" can provide a direct interface to AI processes. Close cooperation is required.[112]

Current status 2022 and further work
The extent to which the contents of SMART standards can be made machine-interpretable depends directly on the extent to which it is possible to capture the structured information required for this already during the standards development process, i.e. within the committee work.

In turn, the type of structuring determines the level of difficulty of this task. This is where the information model comes into play, defining how standards content is fragmented, networked, and metadata is added.

The size of the fragments generated has a significant influence not only on the extent to which the content can be made accessible for reliable automated use, but also on the effort required to create them.

Thus, as the size of the fragments decreases, the importance of user-friendly tool support that minimizes the additional effort required to capture the contents of the standard also increases. This is likely to involve the use of AI-assisted systems that display suggestions for content modelling based on pre-processing, which will be confirmed by the standards authors.

XML documents (NISO-STS) are already being generated in standardization today which have a fragmentation, albeit a coarse one, which is essentially based on the layout structures. However, for systems that are to understand standards content, a corresponding semantic structuring is required.

The theoretical basis is the information model developed in project 2 at CEN/CENELEC, which is currently being further developed at IEC.

It defines the "provision" as a central element and fragment. This is consistent with both the applicable standardization rules (ISO Directives Part 2) and the most important of the use cases identified to date, such as requirements management, and corresponds to Level 3 of the IEC utility model. The aim is to make appropriately fragmented or modelled standard content available to standards users by the end of 2024.

But the next steps to Level 4 of the utility model have already been tested in pilot projects since 2020. In the process, the "provision", the smallest element of the "intermediate" level 3, was further decomposed. The approaches used range from fragmentation using templates into semantic "groups" (e.g., "condition," "subject," "action," "object," etc.) to full modelling of the natural language in semantic triples using RDF.

In order to achieve full machine interpretability of standard content, the experience gained in this process will provide an important basis for the work on SMART standards from 2025 on. This is because only such fully modelled content allows the safe application of all standards content by AI-supported systems.

[112] See Standardization Roadmap AI (1st edition), Chapter 11.4.3 [63].

6 Implementation of the 1st edition of the Standardization Roadmap AI

With the publication of the 1st edition of the Roadmap, the phase of implementation and consolidation of its results began. The aim was to quickly implement as many of the identified recommendations for action as possible with the participation of experts from industry, research and civil society and with the support of the federal ministries. The central objective of this consolidation is to integrate the identified topics in the relevant standardization bodies and to initiate concrete implementation or standardization activities, if possible at European or international level. With the help of the resulting standards and specifications, the identified potential is to be leveraged and the international competitiveness of the German economy supported. The following is the current status of the implementation of the results of the 1st edition.

The first edition of the Standardization Roadmap AI formulates five overarching recommendations for action for seven key topics and a total of 78 needs, some of which vary greatly in character – for example, with regard to the target group or the degree of maturity. To implement the results of the Roadmap, the first step was to develop a consolidation concept aimed at systematically anchoring the needs in the relevant standardization bodies and initiating standardization projects. For this purpose, the 78 needs for action in the Roadmap were analyzed according to their target group and categorized as follows:

→ Category 1: Need addresses standardization
→ Category 2: Need addresses research
→ Category 3: Need addresses policy-makers/legislators
→ Category 4: No need (notes, remarks)

Category 1 comprises recommendations that identify needs for standards and are thus addressed to the standards organizations. Category 2 needs, on the other hand, relate to areas that are the subject of research at this time. The aim here is to initiate research projects and standardization accompanying development at an early stage. The needs of category 3 are primarily addressed to policy-makers and legislators. For example, they aim to revise legal frameworks or regulations, or identify where policy support is needed. Category 4 summarizes notes/remarks or suggestions for procedures that have been taken into account in the preparation of the present Standardization Roadmap.

Figure 51 shows the distribution of the 78 needs of the Standardization Roadmap AI among the four categories.

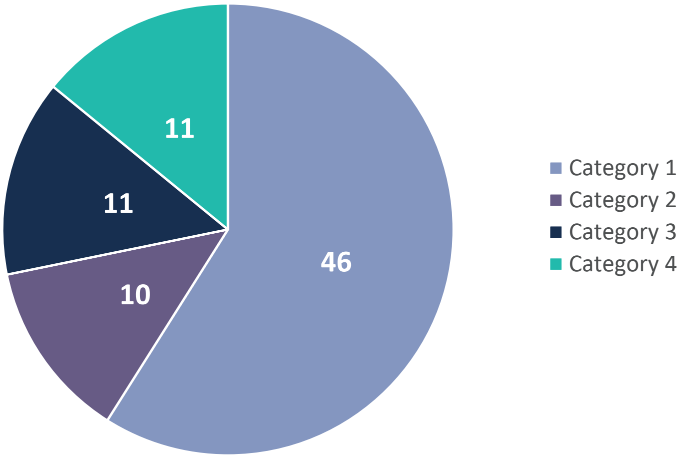


Figure 51: Distribution of needs among categories (As of October 2022, Source: DIN)

6.1 Standardization needs

As expected, the majority of the identified needs are directed at standardization. In order to transfer these Category 1 needs promptly into standardization projects, they were thematically assigned to the relevant standards committees and subjected to further analysis. For the purpose of prioritization, the standardization needs were first assessed according to their degree of maturity (need for concretization or further development) and the urgency of their implementation, and finally discussed in a large number of specialist workshops and meetings with the experts of the committees – always with the aim of integrating the topics in the work programmes of the committees and promptly initiating concrete standardization projects.

Since the required AI expertise is not necessarily available in the committees concerned, the recruitment of new experts for the standardization work represents a critical success factor in the implementation of the requirements. Interested experts are therefore always invited to participate in the relevant committees.

Figure 52 shows the distribution of Category 1 needs among the standards committees. Since a need can often be thematically located in several committees, the simplified presentation shows only those standards committees that are listed as the main points of contact. The figure clearly shows that the Standards Committee Information Technology and selected IT Applications (NIA), in which the DIN/DKE Joint Committee on AI is incorporated, is currently the most relevant standards committee for implementing the requirements.

The various implementation efforts have resulted in a large number of standardization activities, which are described in Figure 53.

Of the 46 identified standardization needs, 20 were integrated into existing standardization projects and 15 new standardization projects were initiated. Table 11 and Table 12 show the standardization needs that were transferred to current standardization projects, and for which new standardization projects were initiated, respectively.

In the case of the remaining eleven requirements in Category 1, it was stated in the discussions with the experts from the standards committees that a transfer to standardization projects is not possible at the present time. The reasons for this are the lack of AI expertise in the relevant standards

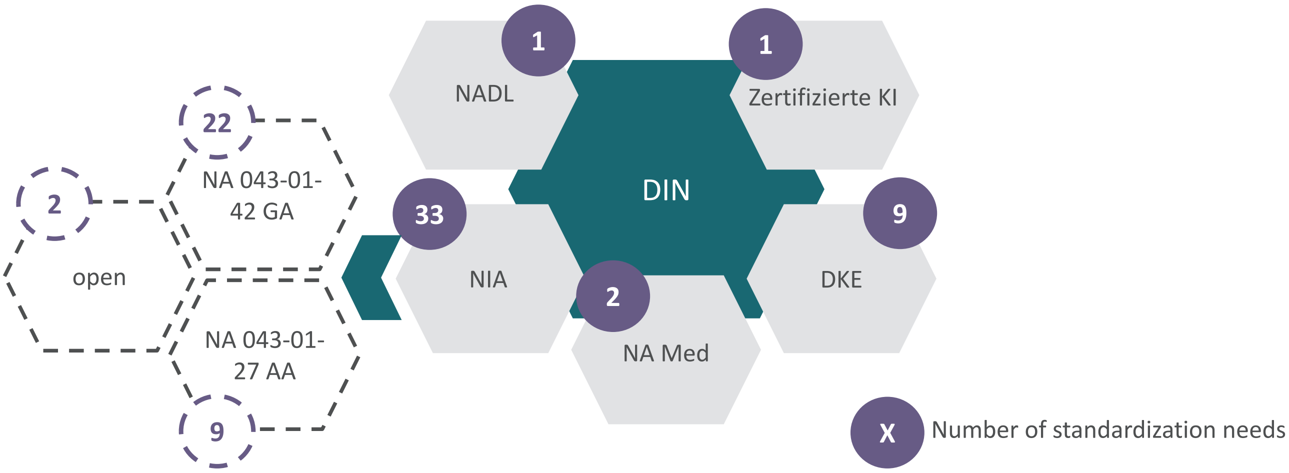


Figure 52: Distribution of standardization needs among the standards committees (As of October 2022, Source: DIN)

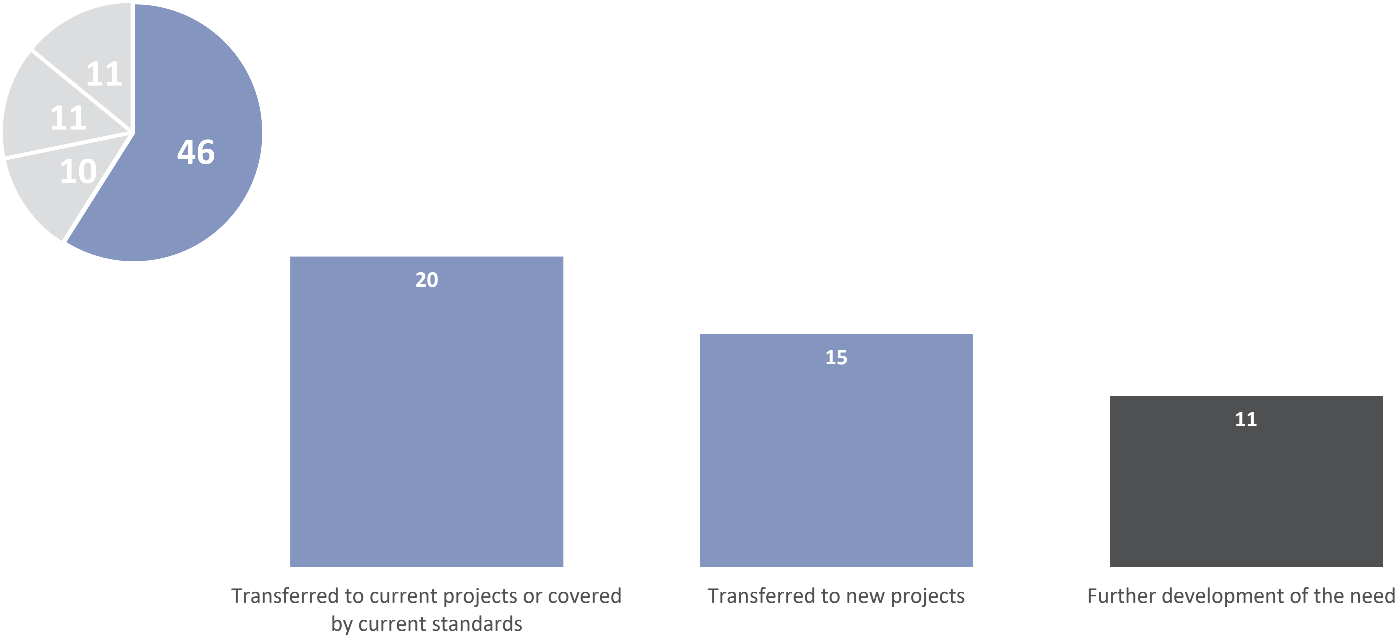


Figure 53: Transfer of requirements to standardization (As of October 2022, Source: DIN)

committees, and the need for further development or concretization of the requirements before they can be passed on to the relevant standards committees and standardization projects initiated.

Table 11: Needs transferred to current standardization projects

Need	Committee	Standard(s)
Define data and its usage	NA 043-01-42 GA	ISO/IEC 2382 [429]
Define controls for IT security	NA 043-01-27-01 AK	DIN EN ISO/IEC 27000 (series) [479]
Risk assessment of IT security with regard to AI systems	NA 043-01-27-01 AK	ISO/IEC 27005:2018 [161]
Standardization of a concept for privacy ethical design	NA 043-01-27-01 AK	DIN EN ISO/IEC 29100:2020 [133], DIN EN ISO/IEC 29134:2020 [134], DIN EN ISO/IEC 27701:2021 [128]
Data quality management for AI	NA 043-01-42 GA	ISO/IEC 5259 (series) [39]
Define the type and quality of data	NA 043-01-42 GA	ISO/IEC 5259 (series) [39]
Design purpose limitation of data	NA 043-01-42 GA, NA 043-01-27-01 AK	ISO/IEC 5259 (series) [39], DIN EN ISO/IEC 27701 [128]
Management system for AI that defines requirements and processes for organizations developing or using AI (taking into account organizational, technical, and process-related test methods as well as test schemes across the entire life cycle of AI systems)	NA 043-01-42 GA	ISO/IEC 42001 [27]
Support of international standardization work on an MSS (management system standard) for AI	NA 043-01-42 GA	ISO/IEC 42001 [27]
Risk management for AI	NA 043-01-42 GA	ISO/IEC 23894:2022 [25]
Re-evaluation of AI systems	NA 043-01-42 GA	ISO/IEC 38507:2022 [26]
Specify restrictions for big data	NA 043-01-42 GA	ISO/IEC TR 20547 (series) [438], [439], [440], [441], [442]
Principles for human-machine-human interaction in the medical sector	NA 063-07-02 AA	DIN EN IEC 81001-5-1:2022-01 – Draft, VDE 0750-103-5-1:2022 [430]
IT security metrics for learning systems and adversarial machine learning (AML)	DIN SPEC 92001-2:2020 WS	DIN SPEC 92001-2:2020 [240]
Criteria for the classification of systems or components within the framework of artificial intelligence	NA 043-01-42 GA	ISO/IEC 5392 [32] ISO/IEC 42001 [27]

Need	Committee	Standard(s)
Collection of terms from different disciplines (glossary)	NA 043-01-41 AA; DKE	ISO/IEC 20924:2021 [431]
Standardized preparation of use cases	NA 043-01-42 GA, IEC TC65/WG23	ISO/IEC TR 24030 [293] PD IEC TR 63283-2 [294] ISO/IEC 22989:2022 [16]
Create data reference model for interoperability	NA 043-01-42 GA	ISO/IEC 20547-3:2020 [440]
Create a function reference model for interoperability	NA 043-01-42 GA	ISO/IEC 42001 [27]
Specify methods for data exchange	NA 043-01-32 AA	ISO/IEC 19763-3:2020 [426]

Table 12: Needs transferred to new standardization projects

Need	Committee	Standard(s)
Test criteria and methods for technical tests of AI solutions	NA 043-01-42 GA	DIN/TS 92004 [427]
Relationship between technical requirements on the one hand and legal and ethical requirements on the other	NA 043-01-42 GA	ISO/IEC TR 29119-11 [132]
Management of transparency and avoidance of discrimination	NA 043-01-42 GA	ISO/IEC TS 12791 [38] ISO/IEC 12792 [238]
Quality backward chain in the AI life cycle	NA 043-01-42 GA	ISO/IEC 5338 [30]
IT security of AI systems in the absence of resource availability (attack vector)	NA 043-01-27-01 AK	ISO/IEC TR 27563 [138]
Design principles for KI systems	NA 043-01-42 GA	ISO/IEC TS 5471 [33], [34] ISO/IEC 5338 [30]
AI security by design and AI security by default	NA 043-01-27-03 AK	ISO/IEC 7699 [428]
IT security criteria for learning systems	NA 043-01-27-03 AK	ISO/IEC 7699 [428]
IT security of training data	NA 043-01-27-03 AK	ISO/IEC 7699 [428]
Criticality levels and IT security	NA 043-01-42 GA	Ad Hoc Group “AI classification” – Draft for CEN/CLC JTC 21 Artificial Intelligence is currently being developed.
IT security criteria for training methods	NA 043-01-42 GA	Ad Hoc Group “AI classification” – Draft for CEN/CLC JTC 21 Artificial Intelligence is currently being developed.

Need	Committee	Standard(s)
Define error classifications, misclassifications and learning from errors	NA 043-01-42 GA	ISO/IEC 42005 [432]
Review process to evaluate existing principles	NA 063-07-02 AA	Work of NA 063-07-02 AA (will be brought to ISO/TC 215 WG 1/2)
Design initial criticality checks of AI systems quickly and easily (risk matrix)	NA 043-01-42 GA	Ad Hoc Group "AI classification" – Draft for CEN/CLC JTC 21 Artificial Intelligence is currently being developed.
Explainable AI		DIN SPEC 92001-3 [117]

6.2 Research needs

Category 2 needs are primarily directed at the research community. The 1st edition of the Roadmap identified ten needs of this type. The aim of the implementation efforts is to initiate research projects. An important pillar of these projects is standardization at an early stage of development, in which the project results are supplied to standardization at an early stage, and thus the transfer of scientific results into marketable products and services is supported quite significantly. Consequently, standardization is a catalyst for innovation that favours the market development, penetration and internationalization of new and further technological developments.

From the results of the 1st edition of the Roadmap, three research projects could be initiated, in which four of the identified research needs are taken up and implemented: ZERTIFIZIERTE KI, AI suitability of standards, and KIMEDS (see Chapter 3.3.1 and Chapter 3.3.2). The remaining six research needs will be given greater detail or further developed within the framework of the present Roadmap.

6.3 Political needs

The 1st edition of the Roadmap formulated a total of eleven needs addressed to policy-makers. To implement these needs, DIN has been actively involved in the political discourse on AI (for example, the AI Act): The results of the Standardization Roadmap were presented to the European Commission and EU parliamentarians, as well as to the German government and members of the Bundestag.

Political demands and recommendations were derived from the identified needs, summarized in a position paper on AI[113], and addressed to policy-makers.

In particular, the position paper calls for links to international standardization and financial support for German experts (especially from small and medium-sized enterprises, science and civil society) to actively participate in international and European standardization projects. Such support is considered essential to ensure that national interests and European values are taken into account.

6.4 Overarching recommendations for action

A total of five overarching recommendations for action were formulated in the 1st edition of the Standardization Roadmap AI. They are of particular importance since they concern all areas of the 1st edition of the Standardization Roadmap AI and are addressed to standardization, research and politics alike.

[113] www.din.de/resource/blob/857886/92863b23027a9737056f-6ca122035931/kurzpositionspapier-kuenstliche-intelligenz-data.pdf

The recommendations for action are currently being implemented – including in research, standardization and implementation projects. The following is the current status (as of: October 2022) of the implementation.

1. **Implement data reference models for the interoperability of AI systems**
   Many different actors come together in value chains. In order for the various AI systems of these actors to be able to work together automatically, a data reference model is needed to exchange data securely, reliably, flexibly, and compatibly. Standards for data reference models from different areas create the basis for a comprehensive data exchange and thus ensure the interoperability of AI systems worldwide.
   Status of implementation:
   Currently, international AI standardization is working intensively on the topic of data. The following ongoing standardization projects should be mentioned in this context:
   - ISO/IEC 5259-1 “Artificial intelligence – Data quality for analytics and machine learning (ML) – Part 1: Overview, terminology, and examples“ [40]
   - ISO/IEC 5259-2 “Artificial intelligence – Data quality for analytics and machine learning (ML) – Part 2: Data quality measures“ [41]
   - ISO/IEC 5259-3 “Artificial intelligence – Data quality for analytics and machine learning (ML) – Part 3: Data quality management requirements and guidelines“ [42]
   - ISO/IEC 5259-4 “Artificial intelligence – Data quality for analytics and machine learning (ML) – Part 4: Data quality process framework“ [43]
   - ISO/IEC 5259-5 “Artificial intelligence – Data quality for analytics and machine learning (ML) – Part 5: Data quality governance“ [44]
   - ISO/IEC 8183 “Information technology – Artificial intelligence – Data life cycle framework“ [45]

   All of these projects are contributing significantly to the implementation of this recommendation for action.
2. **Create a horizontal AI basic security standard**
   AI systems are essentially IT systems – for the latter there are already many standards and specifications from a wide range of application areas. To enable a uniform approach to the IT security of AI applications, an overarching „umbrella standard“ that bundles existing standards and test methods for IT systems and supplements them with AI aspects would be expedient. This basic security standard can then be supplemented by subordinate standards on other topics.
   Status of implementation:
   Currently, intensive work is being done in the AI environment on a common international and European standardization landscape, which should also include the topic of security in the sense of IT security, safety and privacy. At present there are no concrete projects at CEN/CENELEC and ISO/IEC level. To enable a horizontal AI basic security standard, it is imperative to have more AI experts active in standardization, as well as an increased presence in international AI standardization bodies. In particular, persuasive efforts are still needed here.
3. **Design practical initial criticality checks of AI systems**
   When self-learning AI systems decide about people, their possessions or access to scarce resources, unplanned problems in AI can endanger individual fundamental rights or democratic values. So that AI systems in ethically uncritical fields of application can still be freely developed, an initial criticality test should be designed through standards and specifications – this can quickly and legally clarify whether an AI system can even trigger such conflicts.
   Status of implementation:
   In August 2022, the ISO/IEC 42005 [432] project “Information technology – Artificial intelligence – AI system impact assessment” was initiated at the international level under German leadership. This document is a guide for organizations conducting AI system impact assessments for individuals and societies that may be affected by an AI system and its intended and foreseeable uses. It includes considerations on how and when to conduct such assessments and at what stages of an AI system’s life cycle, as well as guidance on documenting impact assessment for AI systems. It also explains how the AI systems impact assessment process can be integrated into an organization’s AI risk management and AI management system. This document is intended for organizations that develop, deploy, or use AI systems. The document is applicable to any organization, regardless of size, type or nature.
4. **Initiate the national implementation programme “Trusted AI” to strengthen the European quality infrastructure**
   So far, there is a lack of reliable quality criteria and test methods for AI systems – this endangers the economic growth and competitiveness of this future technology. A national implementation programme “Trusted AI” is needed, which lays the foundation for reproducible and standardized test methods with which properties of AI systems such as reliability, robustness, performance and functional safety can be tested and statements about

trustworthiness made. Standards and specifications describe requirements for these properties and thus form the basis for the certification and conformity assessment of AI systems. With such an initiative, Germany has the opportunity to develop a certification programme that will be the first of its kind in the world and will be internationally recognized.

Status of implementation:

Various initiatives on "Trusted AI" have been launched in the standardization environment. The focus is on the development of management system standards for the certification of trustworthy handling of AI, as well as the specification of requirements for certifying organizations. Here is a list of the relevant standardization projects:

- ISO/IEC 42001 [27] "Information technology – Artificial intelligence – Management system"
- ISO/IEC 23894 [25] "Information technology – Artificial intelligence – Guidance on risk management"
- ISO/IEC 42005 [432] "Information technology – Artificial intelligence – AI system impact assessment"

In addition, to implement the recommendation for action, the implementation project ZERTIFIZIERTE KI (see Chapter 3.3.2) was launched at the beginning of 2021, in which test criteria, methods and tools for AI systems are to be developed and standardized to enable a comparable assessment of AI systems. A broad participatory process will be used to ensure that methods evolve into generally accepted standards for AI systems and their verification.

5. **Analyze and evaluate use cases for standardization needs**

AI research and the industrial development and application of AI systems are highly dynamic. Already today there are many use cases in the different fields of AI. Standardization needs for AI applications ready for industrial use can be derived from application-typical and industry-relevant use cases. In order to shape standards and specifications, it is important to integrate mutual impulses from research, industry, society and regulation. At the centre of this approach, the developed standards should be tested and further developed on the basis of use cases. In this way, application-specific needs can be identified at an early stage and marketable AI standards realized.

Status of implementation:

The Technical Report ISO/IEC TR 24030:2021 [293] "Information technology – Artificial intelligence (AI) – Use cases" was prepared by the international standards body ISO/IEC/JTC 1/SC 42/WG 4 "Use cases and applications" and published in May 2021. The document contains a collection of use cases and thus provides a good basis for the above-mentioned recommendation for action.

In addition, the Technical Expert Group "Artificial Intelligence Applications in Industrie 4.0 / Intelligent Manufacturing" (TEG AIAI2M) within the German-Chinese Standardization Cooperation Commission (DCKN) also deals with the topic of use cases.

Use case considerations also play an important role at European level within CEN/CENELEC JTC 21. In particular, the activities are related to the planned European regulatory projects (primarily the Data Sovereignity Act and the AI Act), since concrete applications are used there in turn (as examples) for classification, including criticality, and are thus both demonstrative and complementary to the above.

In addition, the implementation project ZERTIFIZIERTE KI (see Chapter 3.3.2) also contributes significantly to the implementation of the recommendation. In sector- and technology-related user groups, participants from industry and science examine concrete industry-specific use cases, always with the aim of defining needs, establishing criteria and benchmarks for testing in practice, and verifying these on the basis of sector-typical use cases. In the next step, the identified needs and findings will be translated into corresponding requirements for a trustworthy use of AI and finally fed into standardization.

6.5 Recruiting experts for standardization

The translation of the identified needs into standardization projects and the subsequent development of standards and specifications is only one objective of the Roadmap's implementation efforts. Another focus is on the recruitment of experts for standardization work. Standardization is a joint task and needs knowledgeable AI experts from industry, science and civil society to actively contribute their knowledge to the development of standards and specifications. Only the early engagement of experts with practical experience and insights will make it possible to develop standards and specifications for AI that are in line with the market and needs. At present, these AI experts are very rarely represented in the standards committees, which makes it much more difficult to quickly translate the requirements into standards.

More than two dozen new experts have been recruited from the consolidation activities to date (as of October 2022), who will henceforth be involved in the standardization committees and contribute their know-how to the development of standards and specifications on AI. This is a good starting point, but it is not sufficient with regard to the diverse potentials and needs identified by the Standardization Roadmap AI. If Germany wants to ensure that its interests are adequately reflected in international AI standards, more AI experts active in standardization are needed and an increased presence in international AI standardization bodies is strongly advised.

6.6 Lighthouse projects

The need for "lighthouse projects of the German Standardization Roadmap AI" was identified by the "Coordination Group AI Standardization and Conformity" as a further consolidation measure. A lighthouse project is understood to be application-typical and industry-relevant use cases that identify standardization requirements for AI-specific applications. With the help of the lighthouse projects, practical experience is to be gathered in the respective application context, concrete needs for standardization are to be derived, and findings on quality and conformity testing are to be obtained. These projects are therefore of particular importance in the implementation of the Standardization Roadmap AI, which is why they enjoy a high level of attention among standardization stakeholders and have great visibility and appeal in industry, research and politics. The concept of lighthouse projects sets out clear framework conditions, for example, on the selection process, evaluation criteria and project sponsorships.

The evaluation criteria take into account, among other things:

→ the balanced participation of all relevant stakeholder groups,
→ the strategic importance and broad impact in macroeconomic terms (e.g. pioneering role, technology leadership),
→ European and/or international connectivity in standardization,
→ regulatory provisions,
→ existing preliminary work from research and implementation projects, and
→ sociotechnical aspects (such as humane work design, organizational conditions, etc.).

In a selection process, the first lighthouse projects or projects with lighthouse character[114] were identified and awarded by the coordination group "AI standardization and conformity", which are described below: safe.trAIn, medical diagnosis and prognosis systems, cloud services, as well as NDE4.0.

Safe.trAIn

The safe.trAIn project[115] (Safe AI using the example of driverless regional trains) is the first official lighthouse project of the AI standardization roadmap. It is funded by the BMWK and since 2022 has been pursuing the goal of linking AI processes with the requirements and approval processes in the rail environment in a practicable manner. The focus of the consortium is on the development of standardized test methods and tools to ensure approval-relevant product safety for a broad use of fully autonomous trains. In addition, the safety architecture is being given detail using the example of the driverless regional train, and a fully automated GoA4 system is being conceptually developed and validated for this use case in a virtual test field. The results of the project are to be transferred into standards and specifications. These play a critical role in accelerating time-to-market and the safe, robust, as well as trustworthy application of AI-based methods for driverless train travel.

AI standards for medical diagnosis and prognosis systems

The application of AI systems in medical diagnostic procedures offers great potential. Even though the number of AI-based medical devices on the market is steadily increasing, the process of developing, manufacturing and launching them on the market, including testing by well-known bodies, has so far been very time-consuming and costly. To increase the use of AI-based medical devices, acceptance and trust must be created on the one hand, and the development and approval processes must be simplified on the other. The aim of the project is therefore to develop standardized test methods and tools for medical AI-based diagnosis and prognosis systems that will enable faster and safer market access. The project, which was recognized by the coordinating group as a "project with lighthouse character", explicitly involves market

[114] A "project with lighthouse character" is characterized by the fact that it sufficiently fulfils the defined evaluation criteria, but its financing and thus also its implementation are not secured at the current moment. Until its implementation, the project will therefore initially be managed as a "project with lighthouse character". With the start of project implementation, it automatically receives the status "Lighthouse Project of the German Standardization Roadmap AI".

[115] https://www.din.de/de/forschung-und-innovation/partner-in-forschungsprojekten/ki/safe-train-860442

participants from industry, regulation, research, and clinical practice to develop marketable AI solutions and increase acceptance and trust for AI-based diagnostic and prognostic systems.

Cloud services

AI solutions are a key technology in digitalization in part because they use scalable cloud technologies. As a result, they remain economical in development and operation, and ensure the international competitiveness of users. The use of platforms, infrastructures and AI frameworks of the large cloud providers basically enables economic market access even for market participants who do not have sufficient IT resources of their own and only little AI expertise. Due to the wide range of applications, the trustworthiness of hybrid and embedded AI solutions is of particular importance, with a large part of the responsibility for the trustworthiness of the technical AI components lying with the cloud providers, developers and operators of the cloud-based AI services. The project was recognized by the coordination group as a "project with lighthouse character".

The key aims of the project are to:

→ make the trustworthiness of AI solutions in the development and operation of cloud-based AI services transparent through internationally recognized conformance testing,

→ develop the test criteria and test methods required for this purpose,

→ introduce the testing principles as a basis for the application-independent horizontal AI standards at the European level, and

→ enable market access to trusted AI at an acceptable cost, even for small and medium-sized enterprises.

NDE 4.0

Nondestructive evaluation (NDE) has always accompanied industrial progress. Germany has played a leading role worldwide in this field for many decades. NDE systems are a central element in the concepts for quality and safety/security technologies in the German economy. This applies in particular to efficient production processes, the safe operation of technical systems and equipment, and end-to-end process methodology. NDE sensor systems have the highest relevance in the context of release, maintenance and servicing processes, traditionally especially in the area of critical or resilient infrastructure (e.g. chemical and plant safety, power generation and distribution, transportation and traffic, construction infrastructure, etc.). With the advancement of digitalization and the advent of AI in NDE, today's standardization process is outdated and established standards no longer cover the rapid developments in NDE. This motivation gave rise to the idea for the project "Innovation acceleration through more flexible validation and certification paths for NDE4.0", which was named as a "project with lighthouse character" by the coordination group. With the help of the project, the process of approval and standardization is to be accelerated, thus providing legal certainty for those affected by the use of NDE4.0.

7 Overview of relevant documents, activities and committees on AI

This chapter serves to provide an overview of already published standards (Chapter 7.1), ongoing standardization activities (Chapter 7.2) and standardization bodies (Chapter 7.3) with relevance for AI. The lists make no claim to completeness.

7.1 Published standards and specifications relevant to AI

Table 13 provides information on already published AI-related standards and specifications, as well as on their relevance for the working groups of the Standardization Roadmap.

Table 13: Overview of published standards and specifications relevant to AI[116]

Document	Title	Date	Body	Relevance for key topics								
				Basic topics	Security/safety	Testing and certification	Sociotechnical systems	Industrial automation	Mobility	Medicine	Financial services	Energy/environment
VDE AR 2842-61 [105]	Development and trustworthiness of autonomous/cognitive systems	2021	DKE/K 801: System Komitee AAL	X	X		X	X		X		X
ISO/IEC 23053 [24]	Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)	2022	NA 043-01-42 GA			X		X		X	X	X
DIN EN 61508-3, VDE 0803-3 [103]	Functional safety of electrical/electronic/programmable electronic safety-related systems – Part 3:Software requirements (IEC 61508-3:2010)	2011	DKE/GK 914 Functional safety of electrical, electronic and programmable electronic systems (E, E, PES) for the protection of persons and the environment		X			X				X
DIN EN 61508-5, VDE 0803-5 [433]	Functional safety of electrical/electronic/programmable electronic safety-related systems – Part 5: Examples of methods for the determination of safety integrity levels (IEC 61508-5:2010)	201102	DKE/GK 914 Functional safety of electrical, electronic and programmable electronic systems (E, E, PES) for the protection of persons and the environment		X	X		X				X
DIN EN 61511-1, VDE 0810-1 [434]	Functional safety – Safety instrumented systems for the process industry sector – Part 1:Framework, definitions, system, hardware and application programming requirements (IEC 61511-1:2016 + COR1:2016 + A1:2017)	201902	DKE/GK 914 Functional safety of electrical, electronic and programmable electronic systems (E, E, PES) for the protection of persons and the environment		X			X				X

[116] This overview makes no claim to completeness.

Document	Title	Date	Body	Relevance for key topics								
				Basic topics	Security/safety	Testing and certification	Sociotechnical systems	Industrial automation	Mobility	Medicine	Financial services	Energy/environment
DIN SPEC 13266 [98]	Guideline for the development of deep learning image recognition systems	2020		X		X		X		X		
DIN EN IEC 62443 (all parts) [435]	Industrial communication networks – Network and system security	2020	DKE/UK 931.1 IT security for industrial automation systems			X		X		X		
ISO/IEC TR 24027 [436]	Information technology – Artificial intelligence (AI) – Bias in AI systems and AI aided decision making	2021	NA 043-01-42 GA	X		X	X	X		X	X	X
ISO/IEC TR 24372 [437]	Information technology – Artificial intelligence (AI) – Overview of computational approaches for AI systems	2021	NA 043-01-42 GA	X		X	X	X		X		X
ISO/IEC TR 24030 [293]	Information technology – Artificial intelligence (AI) – Use cases	2021	NA 043-01-42 GA	X		X		X		X	X	X
ISO/IEC 38507 [26]	Information technology – Governance of IT – Governance implications of the use of artificial intelligence by organizations	2022	NA 043-01-42 GA	X		X	X	X		X		
ISO/IEC TR 24368 [15]	Information technology – Artificial intelligence – Overview of ethical and societal concerns	2022	NA 043-01-42 GA	X			X	X		X	X	X
ISO/IEC TR 24028 [28]	Information technology – Artificial intelligence – Overview of trustworthiness in artificial intelligence	2020	NA 043-01-42 GA	X		X	X	X		X	X	X
ISO/IEC TR 20547-1 [438]	Information technology – Big data reference architecture – Part 1: Framenwork and application process	2020	NA 043-01-42 GA			X						X
ISO/IEC TR 20547-2 [439]	Information technology – Big data reference architecture – Part 2:Use cases and derived requirements	2018	NA 043-01-42 GA			X						X
ISO/IEC 20547-3 [440]	Information technology – Big data reference architecture – Part 3: Reference architecture	2020	NA 043-01-42 GA			X						X
ISO/IEC 20547-4 [441]	Information technology – Big data reference architecture – Part 4: Security and privacy	2020	NA 043-01-42 GA			X	X					X

Document	Title	Date	Body	Relevance for key topics								
				Basic topics	Security/safety	Testing and certification	Sociotechnical systems	Industrial automation	Mobility	Medicine	Financial services	Energy/environment
ISO/IEC TR 20547-5 [442]	Information technology – Big data reference architecture – Part 5: Standards roadmap	2018	NA 043-01-42 GA			X						X
ISO/IEC 20546 [443]	Information technology – Big data – Overview and vocabulary	2019	NA 043-01-42 GA									X
ISO/IEC 33063 [444]	Information technology – Process assessment – Process assessment model for software testing	2015	NA 043-01-07 AA					X		X		
DIN EN ISO/IEC 15408-1 [445]	Information technology – Security techniques – Evaluation criteria for IT security – Part 1: Introduction and general model (ISO/IEC 15408-1:2009)	2020	NA 043-04-27 AA			X		X		X		X
DIN EN ISO/IEC 15408-2 [446]	Information technology – Security techniques – Evaluation criteria for IT security – Part 2: Security functional components (ISO/IEC 15408-2:2008), only on CD-ROM	2020	NA 043-04-27 AA			X		X		X		X
DIN EN ISO/IEC 15408-3 [447]	Information technology – Security techniques – Evaluation criteria for IT security – Part 3: Security assurance components (ISO/IEC 15408-3:2008, Corrected version 2011-06-01), only on CD-ROM	2021	NA 043-04-27 AA			X		X		X		X
ISO/IEC 15408-4 [448]	Information security, cybersecurity and privacy protection – Evaluation criteria for IT security – Part 4: Framework for the specification of evaluation methods and activities	2022	NA 043-04-27 AA			X		X		X		X
ISO/IEC 15408-5 [449]	Information security, cybersecurity and privacy protection – Evaluation criteria for IT security – Part 5: Pre-defined packages of security requirements	2022	NA 043-04-27 AA			X		X				X
DIN EN ISO/IEC 18045 [75]	Information technology – Security techniques – Methodology for IT security evaluation	2021	NA 043-04-27 AA			X		X				X
DIN EN 62304 [353]	Health software – Software life cycle processes	2016	NA 063-01-13 AA		X			X		X		

Document	Title	Date	Body	Relevance for key topics								
				Basic topics	Security/safety	Testing and certification	Sociotechnical systems	Industrial automation	Mobility	Medicine	Financial services	Energy/environment
DIN EN ISO 14971 [351]	Medical devices – Application of risk management to medical devices	2022	NA 063-01-13 AA	X	X			X		X		
ETSI TR 101 583 [450]	Methods for Testing and Specification (MTS); Security Testing; Basic Terminology	2015	European Telecommunications Standards Institute (ETSI)					X		X		X
DIN EN 61513, VDE 0491-2 [451]	Nuclear power plants – Instrumentation and control important to safety – General requirements for systems (IEC 61513:2011)	201309	DKE/UK 967.1 Electrical and control engineering for nuclear facilities		X							X
DIN SPEC 91426 [505]	Quality requirements for video-based methods of personnel selection	2020		X		X	X					
DIN EN 50128; VDE 0831-128 [452]	Railway applications – Communication, signalling and processing systems – Software for railway control and protection systems	201203	UK 351.3 Railway signal systems		X				X			
IEEE 7010 [453]	A New Standard for Assessing the Well-being Implications of Artificial Intelligence	2020	SMC/SC – Standards Committee	X			X			X		
IEEE 2801 [454]	Recommended Practice for the Quality Management of Datasets for Medical Artificial Intelligence	2022	IEEE EMB/Stds Com – Standards Committee					X		X		
DIN ISO 31000 [160]	Risk management – Guidelines (ISO 31000:2018)	2018	NA 175-00-04 AA	X		X	X	X		X	X	X
ISO/SAE 21434 [324]	Road vehicles – Cybersecurity engineering	2021	NA 052-00-32 AA			X						
ISO 26262 series [455]	Road vehicles – Functional safety		NA 052-00-32 AA		X	X						
ISO/TR 4804 [325]	Road vehicles – Safety and cybersecurity for automated driving systems – Design, verification and validation methods	2020	NA 052-00-33-17 AK			X						
DIN EN 62061 [456]	Safety of machinery – Functional safety of safety-related electrical, electronic and programmable electronic control systems	2016	DKE/K 225 Electrotechnical equipment and safety of machinery and mechanical equipment		X	X		X				X
DIN EN ISO 12100 [517]	Safety of machinery – General principles for design – Risk assessment and risk reduction (ISO 12100:2010)	2011	NA 095-01-01 GA		X		X	X				X

Document	Title	Date	Body	Relevance for key topics								
				Basic topics	Security/safety	Testing and certification	Sociotechnical systems	Industrial automation	Mobility	Medicine	Financial services	Energy/environment
DIN CEN ISO/ TR 22100-1 [457]	Safety of machinery – Relationship with ISO 12100 – Part 1: How ISO 12100 relates to type-B and type-C standards	2021	NA 095-01-01 GA		X			X				X
DIN ISO/TR 22100-2, DIN SPEC 33887 [458]	Safety of machinery – Relationship with ISO 12100 – Part 2: How ISO 12100 relates to ISO 13849-1	2014	NA 095-01-01 GA		X			X				X
DIN ISO/TR 22100-3, DIN SPEC 33888 [459]	Safety of machinery – Relationship with ISO 12100 – Part 3: Implementation of ergonomic principles in safety standards	2017	NA 095-01-01 GA		X		X	X				X
DIN CEN ISO/TR 22100-4 [460]	Safety of machinery – Relationship with ISO 12100 – Part 4: Guidance to machinery manufacturers for consideration of related IT-security (cyber security) aspects	2020	NA 095-01-01 GA		X	X		X				X
ISO/TR 22100-5 [461]	Safety of machinery – Relationship with ISO 12100 – Part 5: Implications of artificial intelligence machine learning	2021	NA 095-01-01 GA		X	X	X	X				X
DIN EN ISO 13849-1 [109]	Safety of machinery – Safety-related parts of control systems – Part 1: General principles for design	2016	NA 095-01-03 GA		X	X		X				X
DIN EN ISO 13849-2 [462]	Safety of machinery – Safety-related parts of control systems – Part 2: Validation	2013	NA 095-01-03 GA		X	X		X				X
ISO/IEC 25012 [463]	Software engineering – Software product Quality Requirements and Evaluation (SQuaRE) – Data quality model	2008	NA 043-01-07 AA	X		X		X		X		X
ISO/IEC/IEEE 29119-1 [464]	Software and systems engineering – Software testing – Part 1: General concepts	2022	NA 043-01-07 AA	X		X		X		X		X
ISO/IEC/IEEE 29119-2 [465]	Software and systems engineering – Software testing – Part 2: Test processes	2021	NA 043-01-07 AA	X		X		X		X		X
ISO/IEC/IEEE 29119-3 [466]	Software and systems engineering – Software testing – Part 3: Test documentation	2021	NA 043-01-07 AA	X		X		X		X		X

Document	Title	Date	Body	Relevance for key topics								
				Basic topics	Security/safety	Testing and certification	Sociotechnical systems	Industrial automation	Mobility	Medicine	Financial services	Energy/environment
ISO/IEC/IEEE 29119-4 [467]	Software and systems engineering – Software testing – Part 4: Test techniques	2021	NA 043-01-07 AA	X		X		X		X		X
ISO/IEC/IEEE 29119-5 [468]	Software and systems engineering – Software testing – Part 5: Keyw word-Driven Testing	2016	NA 043-01-07 AA	X						X		X
IEEE 1012 [469]	Standard for System, Software, and Hardware Verification and Validation	2016	IEEE C/S2ESC – Software & Systems Engineering Standards Committee	X		X		X		X		X
IEEE 3333.1.3 [470]	Standard for the Deep Learning-Based Assessment of Visual Experience Based on Human Factors	2022	IEEE C/SAB – Standards Activities Board				X					
ANSI/UL 4600 [471]	Standard for Safety for the Evaluation of Autonomous Products	2022	American National Standards Institute (ANSI)		X			X				X
ISO/IEC/IEEE 12207 [148]	Systems and software engineering – Software life cycle processes	2017	NA 043-01-07 AA	X				X		X		X
ISO/IEC 25000 [472]	Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – Guide to SQuaRE	2014	NA 043-01-07 AA	X		X		X		X		X
ISO/IEC 25024 [473]	Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – Measurement of data quality	2011	NA 043-01-07 AA	X		X		X		X		X
ISO/IEC 25020 [474]	Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – Quality measurement framework	2019	NA 043-01-07 AA	X		X		X		X		X
ISO/IEC 25010 [152]	Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – System and software quality models	2011	NA 043-01-07 AA	X		X		X		X		X
ISO/IEC 25021 [475]	Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – Quality measure elements	2012	NA 043-01-07 AA	X		X		X		X		X

Document	Title	Date	Body	Relevance for key topics								
				Basic topics	Security/safety	Testing and certification	Sociotechnical systems	Industrial automation	Mobility	Medicine	Financial services	Energy/environment
DIN EN ISO 25119-1 [112]	Tractors and machinery for agriculture and forestry – Safety-related parts of control systems	2021	NA 060-16-12 AA		X			X				
DIN SPEC 2343 [476]	Transmission of language-based data between artificial intelligences – Specification of parameters and formats	2020		X								X
ISO/TS 17033 [477]	Ethical claims and supporting information – Principles and requirements	2019	NA 147-00-03 AA	X			X					
DIN EN ISO 26000 [478]	Guidance on social responsibility	2021	NA 175-00-03 AA	X			X					
IEEE 7000 [64]	IEEE Standard Model Process for Addressing Ethical Concerns during System Design	2021	IEEE C/S2ESC – Software & Systems Engineering Standards Committee	X			X			X		
IEEE 7001 [10]	Standard for Transparency of Autonomous Systems	2021	IEEE VT/ITS – Intelligent Transportation Systems	X			X	X				X
IEEE 7002 [11]	Standard for Data Privacy Process	2022	IEEE C/S2ESC – Software & Systems Engineering Standards Committee	X			X	X		X	X	X
IEEE 7007 [12]	Ontological Standard for Ethically driven Robotics and Automation Systems	2021	IEEE RAS/SC – Standing Committee for Standards	X			X			X		
IEEE 7005 [13]	Transparent Employer Data Governance	2021	IEEE C/S2ESC – Software & Systems Engineering Standards Committee	X			X	X				
DIN EN ISO/IEC 27000 [479]	Information technology – Security techniques – Information security management systems – Overview and vocabulary	2020	NA 043-04-27-01 AK				X	X		X		X
DIN EN ISO/IEC 27001 [480]	Information technology – Security techniques – Information security management systems – Requirements	2017	NA 043-04-27-01 AK	X		X		X		X		X
DIN EN ISO/IEC 27002 [481]	Information security, cybersecurity and privacy protection – Information security controls (ISO/IEC 27002:2022)	2017	NA 043-04-27-01 AK			X		X		X		X

Document	Title	Date	Body	Relevance for key topics								
				Basic topics	Security/safety	Testing and certification	Sociotechnical systems	Industrial automation	Mobility	Medicine	Financial services	Energy/environment
DIN EN ISO/IEC 27701 [128]	Security techniques – Extension to ISO/IEC 27001 and ISO/IEC 27002 for privacy information management – Requirements and guidelines	2021	NA 043-04-27-05 AK	X		X		X		X		X
DIN EN ISO/IEC 17000 [147]	Conformity assessment	2020	NA 147-00-03 AA	X		X	X	X		X		X
ITU-T Y.qos-ml-arc [482]	Architecture of machine learning based QoS assurance for the IMT-2020 network	2017	ITU-T SG 13 – Future networks	X								X
ETSI TS 103 195-2 [483]	Autonomic network engineering for the self-managing Future Internet (AFI); Generic Autonomic Network Architecture; Part 2: An Architectural Reference Model for Autonomic Networking, Cognitive Networking and Self-Management	2018	ETSI Autonomic network engineering for the self-managing Future Internet (AFI)									X
DIN EN ISO/IEC 17011 [159]	Conformity assessment – Requirements for accreditation bodies accrediting conformity assessment bodies	2018	NA 147-00-03 AA	X		X						
DIN EN ISO/IEC 17020 [157]	Conformity assessment – Requirements for the operation of various types of bodies performing inspection	2012	NA 147-00-03 AA	X		X						
DIN EN ISO/IEC 17021-1 [22]	Conformity assessment – Requirements for bodies providing audit and certification of management systems – Part 1: Requirements	2015	NA 147-00-03 AA	X		X						
DIN EN ISO/IEC 17021-2 [484]	Conformity assessment – Requirements for bodies providing audit and certification of management systems – Part 2: Competence requirements for auditing and certification of environmental management systems	2019	NA 147-00-03 AA	X		X						
DIN EN ISO/IEC 17021-3 [485]	Conformity assessment – Requirements for bodies providing audit and certification of management systems – Part 3: Competence requirements for auditing and certification of quality management systems	2019	NA 147-00-03 AA	X		X						

Document	Title	Date	Body	Relevance for key topics								
				Basic topics	Security/safety	Testing and certification	Sociotechnical systems	Industrial automation	Mobility	Medicine	Financial services	Energy/environment
DIN EN ISO/IEC 7024 [155]	Conformity assessment – General requirements for bodies operating certification of persons	2012	NA 147-00-03 AA	X		X						
DIN EN ISO/IEC 17025 [156]	General requirements for the competence of testing and calibration laboratories	2018	NA 147-00-03 AA	X		X		X				
DIN EN ISO/IEC 17029 [158]	Conformity Assessment – General principles and requirements for validation and verification bodies	2020	NA 147-00-03 AA	X		X						
DIN EN ISO/IEC 17030 [486]	Conformity assessment – General requirements for third-party marks of conformity	2021	NA 147-00-03 AA	X		X						
DIN EN ISO/IEC 17040 [487]	Conformity assessment – General requirements for peer assessment of conformity assessment bodies and accreditation bodies	2005	NA 147-00-03 AA	X		X						
DIN EN ISO/IEC 17043 [488]	Conformity assessment – General requirements for the competence of proficiency testing providers	2022	NA 147-00-03 AA	X		X						
DIN EN ISO/IEC 17050-1 [489]	Conformity assessment – Supplier's declaration of conformity – Part 1: General requirements	2010	NA 147-00-03 AA	X		X		X				
DIN EN ISO/IEC 17050-2 [490]	Conformity assessment – Supplier's declaration of conformity – Part 2: Supporting documentation	2005	NA 147-00-03 AA	X		X		X				
DIN EN ISO/IEC 17065 [17]	Conformity assessment – Requirements for bodies certifying products, processes and services	2013	NA 147-00-03 AA	X		X						
DIN EN ISO/IEC 17067 [18]	Conformity assessment – Fundamentals of product certification and guidelines for product certification schemes	2013	NA 147-00-03 AA	X		X		X		X		
ITU-T F.AI-DLFE [491]	Deep Learning Software Framework Evaluation Methodology	2021	ITU-T SG 16 – Multimedia	X						X		X
ITU-T Y.3173 [492]	Framework for evaluating intelligence level of future networks including IMT-2020	2020	ITU-T SG 13 – Future networks									X

Document	Title	Date	Body	Relevance for key topics								
				Basic topics	Security/safety	Testing and certification	Sociotechnical systems	Industrial automation	Mobility	Medicine	Financial services	Energy/environment
ISO/IEC 27034-1 [122]	Information technology – Security techniques – Application security – Part 1: Overview and concepts	2011	NA 043-04-27 AA			X		X				X
ISO/IEC 27034-2 [123]	Information technology – Security techniques – Application security – Part 2: Organization normative framework	2015	NA 043-04-27 AA			X		X				X
ISO/IEC 27034-3 [124]	Information technology – Application security – Part 3: Application security management process	2018	NA 043-04-27 AA			X		X				X
ISO/IEC 27034-5 [125]	Information technology – Security techniques – Application security – Part 5: Protocols and application security controls data structure	2017	NA 043-04-27 AA			X		X				X
ISO/IEC 27034-6 [126]	Information technology – Security techniques – Application security – Part 6: Case studies	2016	NA 043-04-27 AA			X						X
ISO/IEC 27034-7 [127]	Information technology – Security techniques – Application security – Part 7: Assurance prediction framework	2018	NA 043-04-27 AA			X		X				X
DIN EN ISO/IEC 29101 [493]	Information technology – Security techniques – Privacy architecture framework	2022	NA 043-04-27 AA			X		X				X
DIN EN ISO/IEC 29134 [134]	Information technology – Security techniques – Guidelines for privacy impact assessment	2020	NA 043-04-27 AA			X		X				X
DIN EN ISO/IEC 29147 [494]	Information technology – Security techniques – Vulnerability disclosure	2020	NA 043-04-27 AA			X		X				X
DIN EN ISO/IEC 29151 [135]	Information technology – Security techniques – Code of practice for personally identifiable information protection	2022	NA 043-04-13 GA			X		X				X
DIN EN ISO/IEC 29100 [133]	Information technology – Security techniques – Privacy framework	2020	NA 043-04-27 AA					X		X		X
ITU-T F.AI-DLPB [495]	Metrics and evaluation methods for deep neural network processor benchmark	2020	ITU-T SG 16 – Multimedia							X		X

Document	Title	Date	Body	Relevance for key topics								
				Basic topics	Security/safety	Testing and certification	Sociotechnical systems	Industrial automation	Mobility	Medicine	Financial services	Energy/environment
ITU-T Y.3170 [496]	Requirements for machine learning – based quality of service assurance for the IMT-2020 Network	2018	ITU-T SG 13 – Future networks				X					X
ETSI DGR SAI 002 [497]	Securing Artificial Intelligence (SAI); Data Supply Chain Report	2021	ETSI “Securing Artificial Intelligence (SAI)“			X		X		X	X	X
ETSI DGS SAI 003 [336]	Securing Artificial Intelligence (SAI); Security Testing of AI	2022	ETSI “Securing Artificial Intelligence (SAI)“			X		X		X	X	X
ETSI TS 103 296 [498]	Speech and Multimedia Transmission Quality (STQ); Requirements for Emotion Detectors used for Telecommunication Measurement Applications; Detectors for written text and spoken speech	2016	ETSI “Speech and Multimedia Transmission Quality (STQ)“				X					
ETSI GR ENI 004 [499]	Experiential Networked Intelligence (ENI); Terminology for Main Concepts in ENI Disclaimer	2019	ETSI “Experiential Networked Intelligence (ENI)“							X		
ISO/TR 24291 [501]	Health informatics – Applications of machine learning technologies in imaging and other medical applications	2021	ISO TC 215					X		X		
ISO/TR 3985 [502]	Biotechnology – Data publication – Preliminary considerations and concepts	2021	ISO TC 276							X		
ISO/TS 22756 [503]	Health Informatics – Requirements for a knowledge base for clinical decision support systems to be used in medication-related processes	2020	ISO TC 215							X		
DIN SPEC 92001-1 [162]	Artificial Intelligence – Life Cycle Processes and Quality Requirements – Part 1: Quality Meta Model	2019	DIN SPEC Consortium	X			X	X		X	X	
DIN SPEC 92001-2 [240]	Artificial Intelligence – Life Cycle Processes and Quality Requirements – Part 2: Robustness	2020	DIN SPEC Consortium	X		X	X	X		X		X
DIN SPEC 13288 [506]	Guideline for the development of deep learning image recognition systems in medicine	2021	DIN SPEC Consortium					X		X		

Document	Title	Date	Body	Relevance for key topics								
				Basic topics	Security/safety	Testing and certification	Sociotechnical systems	Industrial automation	Mobility	Medicine	Financial services	Energy/environment
ISO/TS 5346 [507]	Health informatics – Categorial structure for representation of traditional Chinese medicine clinical decision support system	2022	ISO/TC 215							X		
Series DIN EN ISO 11073 [508]	Health informatics – Personal health device communication		ISO/TC 215							X		
DIN CEN ISO/TS 22703 [509]	Health informatics – Requirements for medication safety alerts (ISO/TS 22703:2021)	2022	ISO/TC 215		X					X		
ISO/TR 19669 [510]	Health informatics – Re-usable component strategy for use case development	2017	ISO/TC 215							X		
IEEE P2802 [511]	Standard for the Performance and Safety Evaluation of Artificial Intelligence Based Medical Device: Terminology	2022	IEEE AIMDWG – Artificial Intelligence Medical Device Working Group				X	X		X		X
DIN EN ISO 13485 [381]	Medical devices – Quality management systems – Requirements for regulatory purposes (ISO 13485:2016)	2021	NA 063-01-13 AA							X		
DIN EN 62366-1 [355]	Medical devices – Part 1: Application of usability engineering to medical devices (IEC 62366-1:2015 + COR1:2016 + A1:2020)	2021	UK 811.4		X					X		
DIN EN 82304-1 [354]	Health Software – Part 1: General requirements for product safety	2018	DKE/UK 811.3		X					X		
DIN EN 60601-1-10 [375]	Medical electrical equipment – Part 1-10: General requirements for basic safety and essential performance – Collateral Standard: Requirements for the development of physiologic closed-loop controllers (IEC 60601-1-10:2007 + A1:2013 + A2:2020)	2021	DKE/K 811		X					X		
IEC/TR 60601-4-1 [373]	Guidance and interpretation – Medical electrical equipment and medical electrical systems employing a degree of autonomy	2017	TC 62/SC 62A							X		
ISO/TR 24971 [352]	Medical devices – Guidance on the application of ISO 14971:2022	2020	ISO/TC 210							X		

Document	Title	Date	Body	Relevance for key topics								
				Basic topics	Security/safety	Testing and certification	Sociotechnical systems	Industrial automation	Mobility	Medicine	Financial services	Energy/environment
IEC/TR 62366-2 [357]	Medical devices – Part 2: Guidance on the application of usability engineering to medical devices	2021	ISO/TC 210							X		
DIN EN 62267, VDE 0831-267 [332]	Railway applications – Automated Urban Guided Transport (AUGT) – Safety requirements (IEC 62267:2009)	2010	DKE/UK 351.3		X				X			
DIN VDE V 0831-103 [343]	Electric signalling systems for railways – Part 103: Identification of safety requirements for technical functions in railway signalling	2020	DIN and VDE		X				X			
DIN VDE V 0831-101 [344]	Electric signalling systems for railways – Part 101: Semi-quantitative processes for risk analysis of technical functions in railway signalling	2022	DIN and VDE		X				X			
ISO 22737 [327]	Intelligent transport systems – Low-speed automated driving (LSAD) systems for predefined routes – Performance requirements, system requirements and performance test procedures	2021	ISO/TC 204		X				X			
VDE SPEC 90012 [242]	VCIO based description of systems for AI trustworthiness characterisation	2022					X			X		
VDI-MT 7001 [512]	Communication and public participation in construction and infrastructure projects – Standards for work stages of engineers	2021					X					
DIN EN ISO 26800 [239]	Ergonomics – General approach, principles and concepts	2011					X					
DIN EN ISO 6385 [235]	Ergonomics principles in the design of work systems	2016					X					
DIN EN ISO 10075 (all parts) [513]	Ergonomic principles related to mental workload						X					
DIN EN ISO 11064 [243]	Ergonomic design of control centres	2011					X					
DIN EN ISO 9241 (all parts) [514]	Ergonomic requirements for office work with visual display terminals (VDTs)						X					

Document	Title	Date	Body	Relevance for key topics								
				Basic topics	Security/safety	Testing and certification	Sociotechnical systems	Industrial automation	Mobility	Medicine	Financial services	Energy/environment
DIN EN 614-1 [180]	Safety of machinery – Ergonomic design principles – Part 1: Terminology and general principles;	2009			X		X					
DIN EN 614-2 [181]	Safety of machinery – Ergonomic design principles – Part 2: Interactions between the design of machinery and work tasks	2008			X		X					
DIN EN 894 (all parts) [515]	Safety of machinery – Ergonomics requirements for the design of displays and control actuators				X		X					
DIN EN 16710-2 [516]	Ergonomics methods – Part 2: A methodology for work analysis to support design	2016					X					
ISO/TR 16982 [518]	Ergonomics of human-system interaction – Usability methods supporting human-centred design	2002					X					
DIN EN ISO 27500 [271]	The human-centred organization – Rationale and general principles	2017					X					
VDI/VDE-MT 7100 [241]	Learning-friendly work design – Goals, benefits, terms and definitions	2022					X					
DIN EN 15804 [413]	Sustainability of construction works – Environmental product declarations – Core rules for the product category of construction products	2022	NA 005-01-31 AA									X
DIN EN ISO 14044 [412]	Environmental management – Life cycle assessment – Requirements and guidelines (ISO 14044:2006 + Amd 1:2017 + Amd 2:2020)	2021	NA 172-00-03-AA									X
DIN EN ISO 14040 [411]	Environmental management – Life cycle assessment – Principles and framework (ISO 14040:2006 + Amd 1:2020)	2021	NA 172-00-03-AA									X
DIN EN ISO 14026 [410]	Environmental labels and declarations – Principles, requirements and guidelines for communication of footprint information (ISO 14026:2017)	2018	NA 172-00-03-AA									X

Document	Title	Date	Body	Relevance for key topics								
				Basic topics	Security/safety	Testing and certification	Sociotechnical systems	Industrial automation	Mobility	Medicine	Financial services	Energy/environment
ISO 21930 [519]	Sustainability in buildings and civil engineering works – Core rules for environmental product declarations of construction products and services	2017	ISO/TC 59, Building and civil engineering works, Subcommittee SC 17, Sustainability in buildings and civil engineering works									X
ISO/TS 14048 [521]	Environmental Management – Life Cycle Assessment – Data Documentation Format	2002	ISO/TC 207, Environmental management, Subcommittee SC 5, Life cycle assessment									X
CWA 17284 [522]	Materials modelling – Terminology, classification and metadata	2018	CEN/CENELEC, WS									X
CWA 17815 [523]	Materials characterisation – Terminology, metadata and classification	2021	CEN/CENELEC, WS									X
DIN IEC/ TS 62998-1:2021-10, VDE V 0113-998-1 [520]	Safety of machinery – Safety-related sensors used for the protection of persons (IEC TS 62998-1:2019)	2021	IEC/TC 44 Safety of machinery – Electrotechnical aspects		X							
ISO/IEC 22989 [16]	Artificial intelligence – Concepts and terminology	2022-07	NA 043-01-42 GA	X	X	X	X	X	X	X		X
ISO/IEC 23894 [25]	Information Technology – Artificial Intelligence – Risk Management	2022	NA 043-01-42 GA	X		X	X	X	X	X	X	X
ISO/IEC 19763-3 [426]	Information technology – Metamodel framework for interoperability (MFI) – Part 3: Metamodel for ontology registration	2020	NA 043-01-32 AA	X								

7.2 Current standardization activities with relevance for AI

Table 14 lists a selection of current activities on the topic of AI. The table does not claim to be complete.

Table 14: Overview of current standardization activities with relevance for AI

Document	Title	Short description	Body	Relevance for key topics								
				Basic topics	Security/safety	Testing and certification	Sociotechnical systems	Industrial automation	Mobility	Medicine	Financial services	Energy/environment
IEEE P2846	A Formal Model for Safety Considerations in Automated Vehicle Decision Making	Technology-neutral mathematical model and test method for automated decision-making regarding vehicles	IEEE VT/ITS – Intelligent Transportation Systems		X			X	X			
ISO/IEC 5259-2	Artificial intelligence – Data quality for analytics and machine learning (ML) – Part 2: Data quality measures		NA 043-01-42 GA	X		X		X		X	X	X
ISO/IEC 5259-5	Artificial intelligence – Data quality for analytics and machine learning (ML) – Part 5: Data quality governance	This document provides a data quality governance framework for analytics and machine learning to enable governing bodies of organizations to direct and oversee the implementation and operation of data quality measures, management, and related processes with adequate controls throughout the data life cycle. This document can be applied to any analytics and machine learning. This document does not define specific management requirements or process requirements specified in 5259-3 and 5259-4 respectively.	NA 043-01-42 GA	X		X	X	X		X	X	X
ISO/IEC TR 5469	Artificial intelligence – Functional safety and AI systems	The document is intended to describe characteristics, relevant risk factors, usable methods, and processes for the application of AI in safety-related functions to control AI systems and for the application of AI in the development of safety-related functions. It will be developed in collaboration with IEC SC65A (the standardization group responsible for IEC 61508).	NA 043-01-42 GA		X	X	X	X	X	X	X	X

Document	Title	Short description	Body	Relevance for key topics								
				Basic topics	Security/safety	Testing and certification	Sociotechnical systems	Industrial automation	Mobility	Medicine	Financial services	Energy/environment
ISO/IEC TS 5471	Artificial intelligence – Quality evaluation guidelines for AI systems		NA 043-01-42 GA	X		X	X	X		X		X
ISO/IEC 24029-2	Artificial intelligence (AI) – Assessment of the robustness of neural networks – Part 2: Methodology for the use of formal methods		NA 043-01-42 GA	X		X		X	X	X	X	X
ISO/IEC TR 24029-1	Artificial Intelligence (AI) – Assessment of the robustness of neural networks – Part 1: Overview	This standard looks at the robustness of AI systems and provides an overview of the approaches and methods available for assessing problems and risks related to robustness. A particular focus is on neural networks, their functionality and usability.	NA 043-01-42 GA	X		X		X	X	X	X	X
ISO/IEC 5259-1	Data quality for analytics and ML – Part 1: Overview, terminology, and examples	Data quality management for machine learning: Overview, terminology and examples	NA 043-01-42 GA	X		X	X	X	X	X	X	X
ISO/IEC 5259-3	Data quality for analytics and ML – Part 3: Data Quality Management Requirements and Guidelines	Data quality management for machine learning: Requirements and guidelines	NA 043-01-42 GA	X		X	X	X	X	X	X	X
ISO/IEC 5259-4	Data quality for analytics and ML – Part 4: Data quality process framework	Data quality management for machine learning: Processes	NA 043-01-42 GA	X		X	X	X	X	X	X	X

Document	Title	Short description	Body	Relevance for key topics								
				Basic topics	Security/safety	Testing and certification	Sociotechnical systems	Industrial automation	Mobility	Medicine	Financial services	Energy/environment
ISO/IEC TS 8200	Information technology – Artificial intelligence – Controllability of automated artificial intelligence systems	This document defines a basic framework with principles, characteristics and approaches for the realization and enhancement for automated artificial intelligence (AI) systems controllability. The following areas are covered: — State observability and state transition — Control transfer process and cost — Reaction to uncertainty during control transfer — Verification and validation approaches This document is applicable to all types of organizations (e. g. commercial enterprises, government agencies, not-for-profit organizations) developing and using AI systems during their whole life cycle.	NA 043-01-42 GA	X		X	X	X		X		X
ISO/IEC 8183	Information technology – Artificial intelligence – Data life cycle framework	This document provides an overarching data life cycle framework that is instantiable for any AI system from data ideation to decommission. This document is applicable to the data processing throughout the AI system life cycle including the acquisition, creation, development, deployment, maintenance and decommissioning. This document does not define specific services, platforms or tools. This document is applicable to all organizations, regardless of type, sizes and nature, that use data in the development and use of AI systems.	NA 043-01-42 GA	X		X		X		X	X	X
ISO/IEC 42001	Information Technology – Artificial intelligence – Management system		NA 043-01-42 GA	X		X	X	X		X		X

Document	Title	Short description	Body	Relevance for key topics								
				Basic topics	Security/safety	Testing and certification	Sociotechnical systems	Industrial automation	Mobility	Medicine	Financial services	Energy/environment
ISO/IEC TS 6254	Information technology – Artificial intelligence – Objectives and approaches for explainability of ML models and AI systems	This document describes approaches and methods that can be used to achieve explainability objectives of stakeholders with regards to ML models and AI systems' behaviours, outputs, and results. Stakeholders include but are not limited to, academia, industry, policy makers, and end users. It provides guidance concerning the applicability of the described approaches and methods to the identified objectives throughout the AI system's life cycle, as defined in ISO/IEC 22989:2022.	NA 043-01-42 GA	X		X	X	X		X		X
ISO/IEC TR 29119-11	Information technology – Artificial intelligence – Testing for AI systems – Part 11:	This document describes testing techniques (including those described in ISO/IEC/IEEE 29119-4:2021) applicable for AI systems in the context of the AI system life cycle model stages defined in ISO/IEC 22989:2022. It describes how AI and ML assessment metrics can be used in the context of those testing techniques. It also maps testing processes, including those described in ISO/IEC/IEEE 29119-2:2021, to the verification and validation stages in the AI system life cycle.	NA 043-01-42 GA	X		X	X	X		X		X
ISO/IEC 12792	Information technology – Artificial intelligence – Transparency taxonomy of AI systems	This document defines a taxonomy of information elements to assist AI stakeholders with identifying and addressing the needs for transparency of AI systems. The document describes the semantics of the information elements and their relevance to the various objectives of different AI stakeholders. This document uses a horizontal approach and is applicable to any kind of organization and application involving AI. V02/	NA 043-01-42 GA	X			X	X		X	X	X

Document	Title	Short description	Body	Relevance for key topics								
				Basic topics	Security/safety	Testing and certification	Sociotechnical systems	Industrial automation	Mobility	Medicine	Financial services	Energy/environment
ISO/IEC TS 12791	Information technology – Artificial intelligence – Treatment of unwanted bias in classification and regression machine learning tasks	This document provides mitigation techniques that can be applied throughout the AI system life cycle in order to treat unwanted bias. This document describes how to address unwanted bias in AI systems that use machine learning to conduct classification and regression tasks. This document is applicable to all types and sizes of organization.	NA 043-01-42 GA	X		X	X			X		X
ISO/IEC FDIS 24668	Information technology – Artificial intelligence – Process management framework for Big data analytics	Management für Datenanalysen im Bereich Big Data	NA 043-01-42 GA				X				X	X
ISO/IEC 5338	Information technology – Artificial intelligence – AI system life cycle processes	Terminology standard on life cycle processes of AI systems (voting phase)	NA 043-01-42 GA	X		X		X		X	X	X
ISO/IEC TS 4213	Information technology – Artificial Intelligence – Assessment of machine learning classification performance	Metrics for the performance capability of AI	NA 043-01-42 GA	X		X	X	X	X	X		X
ISO/IEC 5339	Information Technology – Artificial Intelligence – Guidelines for AI Applications	Guidelines for application of AI systems (in voting phase)	NA 043-01-42 GA	X			X	X		X		X
ISO/IEC 5394	Information Technology – Artificial intelligence – Management System	Management system standard for AI	ISO/IEC JTC 1/SC 32			X	X	X		X		X
ISO/IEC 5392	Information technology – Artificial intelligence – Reference Architecture of Knowledge Engineering	Reference architecture for knowledge-based systems	NA 043-01-42 GA	X		X	X	X		X	X	X

Document	Title	Short description	Body	Relevance for key topics								
				Basic topics	Security/safety	Testing and certification	Sociotechnical systems	Industrial automation	Mobility	Medicine	Financial services	Energy/environment
ISO/IEC TS 24462	Ontology for ICT Trustworthiness Assessment	New project for a Technical Specification. Developed in ISO/IEC JTC 1/ WG 13 "Trustworthiness".	ISO/IEC JTC 1/SC 27	X			X	X		X		X
ISO 24089	Road vehicles – Software update engineering	New standard in development	ISO/TC 22/SC 32						X			
ISO/IEC 25059	Software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) Quality Model for AI-based systems	Quality assessment for AI-based systems	NA 043-01-42 GA	X		X	X	X	X	X		X
IEEE P7003	Algorithmic Bias Considerations		IEEE C/S2ESC – Software & Systems Engineering Standards Committee	X			X			X		X
IEEE P7006	Standard on Personal Data AI Agent Working Group			X			X					
IEEE P7008	Standard for Ethically Driven Nudging for Robotic, Intelligent and Autonomous Systems		IEEE RAS/SC – Standing Committee for Standards	X			X	X		X		X
IEEE P7009	Standard for Fail-Safe Design of Autonomous and Semi-Autonomous Systems		IEEE RAS/SC – Standing Committee for Standards	X			X	X				X
IEEE P7011	Standard for the Process of Identifying & Rating the Trustworthiness of News Sources		IEEE SSIT/SC – Social Implications of Technology Standards Committee	X			X	X				X
IEEE P7012	Standard for Machine Readable Personal Privacy Terms		IEEE SSIT/SC – Social Implications of Technology Standards Committee	X			X			X		X

Document	Title	Short description	Body	Relevance for key topics								
				Basic topics	Security/safety	Testing and certification	Sociotechnical systems	Industrial automation	Mobility	Medicine	Financial services	Energy/environment
IEEE P7014	Standard for Ethical considerations in Emulated Empathy in Autonomous and Intelligent Systems		IEEE SSIT/SC – Social Implications of Technology Standards Committee	X			X			X		
NISTIR 8269	A Taxonomy and Terminology of Adversarial Machine Learning	The taxonomy orders different types of attacks, defenses and consequences. Terminology defines key terms related to the security of ML in AI systems.		X						X	X	X
ISO/IEC 27005	Information security, cybersecurity and privacy protection – Guidance on managing information security risks		NA 043-04-27-01 AK				X	X		X	X	X
ETSI DTR INT 008 (TR 103 821)	Autonomic network engineering for the self-managing Future Internet (AFI); Artificial Intelligence (AI) in Test Systems and Testing AI models.	Test framework for network automation systems such as ETSI GANA (Generic Autonomic Networking Architecture)										X
ISO/IEC TR 17866	Artificial intelligence – Best practice guidance for mitigating ethical and societal concerns		NA 043-01-42 GA	X						X		
ISO/IEC 42005	Information technology – Artificial intelligence – AI system impact assessment		NA 043-01-42 GA	X				X		X		
ISO/IEC NP TS 17847	Information technology – Artificial intelligence – Verification and validation analysis of AI systems		NA 043-01-42 GA	X				X		X		

Document	Title	Short description	Body	Relevance for key topics								
				Basic topics	Security/safety	Testing and certification	Sociotechnical systems	Industrial automation	Mobility	Medicine	Financial services	Energy/environment
ISO/IEC TR 17903	Information technology – Artificial intelligence – Overview of machine learning computing devices		NA 043-01-42 GA					X		X		
ISO TS 23543	Guidance for developing cybersecurity requirements in anaesthetic and respiratory equipment standards	This document is intended to provide guidance for the application of cybersecurity in safety standards for anaesthetic and respiratory equipment. It is intended to assist each committee in identifying, assessing, and addressing cybersecurity risks, and in the preparation of corresponding requirements in an appropriate and consistent way. This document is applicable to particular device standards for anaesthetic and respiratory equipment with external (accessible) data interfaces (Signal Input/Output Part (SIP/SOP)).	ISO TC 121							X		
DIN SPEC 92001-3	Artificial Intelligence_- Life Cycle Processes and Quality Requirements_- Part_3: Explainability	Sector-independent guide to suitable approaches and methodologies for promoting explainability throughout the life cycle of an AI model	DIN SPEC Consortium	X				X		X		

Document	Title	Short description	Body	Relevance for key topics								
				Basic topics	Security/safety	Testing and certification	Sociotechnical systems	Industrial automation	Mobility	Medicine	Financial services	Energy/environment
ISO/TS 9491	Biotechnology – Recommendations and requirements for predictive computational models in personalized medicine research – Part 1: Guidelines for constructing, verifying and validating models	This document defines challenges and requirements for predictive computational models constructed for research purposes in personalized medicine. It specifies recommendations and requirements for the setup, formatting, validation, simulation, storing and sharing of such models, as well as their application in clinical trials and other research areas. It summarizes specific challenges regarding data input, as well as verifying and validating of such models that can be considered as best practices for modelling in research and development in the field of personalized medicine. This document also specifies recommendations and requirements for data used to construct or needed for validating models, including rules and requirements for formatting, description, annotation, interoperability, integration, accessing, as well as recording and documenting the provenance of such data. This document does not provide specific rules or requirements for the use of computational models in the clinical routine, or for diagnostic or therapeutic purposes.	ISO/TC 276							X		

Document	Title	Short description	Body	Relevance for key topics								
				Basic topics	Security/safety	Testing and certification	Sociotechnical systems	Industrial automation	Mobility	Medicine	Financial services	Energy/environment
PT 63450	Artificial Intelligence-enabled Medical Devices – Methods for the Technical Verification and Validation	This document establishes methods for medical device manufacturers to verify and validate artificial intelligence / machine learning-enabled medical devices (AI/ML-MD), i. e. medical devices that use artificial intelligence, in part or in whole, to achieve their intended medical purpose. This includes verification and validation activities for the model of the artificial intelligence as well as selection, metrological characterization and management of the datasets. Such activities are implemented at various stages of the medical device life cycle, especially including design control, monitoring and design change. This document is also applicable to any hardware or software utilizing artificial intelligence that impacts the intended use of a medical device.	IEC/TC 62						X	X		
ISO PAS 8800	Road vehicles – Safety and AI	This document defines safety-related characteristics and risk factors that impact artificial intelligence (AI) underperformance and faulty behaviour in a road vehicle context. It describes a framework that takes into account all phases of the development and deployment life cycle. This includes deriving suitable functional safety requirements, considering data quality and completeness, architectural measures to control and mitigate errors, tools to support AI, verification and validation techniques, and the evidence needed to ensure the overall safety of the system.	ISO/TC 22/SC 32		X				X	X		
ISO/DIS 34501	Road vehicles – Terms and definitions of test scenarios for automated driving systems	Defines basic terms related to scenarios and scenario-based testing	ISO/TC 22/SC 33						X	X		

Document	Title	Short description	Body	Relevance for key topics								
				Basic topics	Security/safety	Testing and certification	Sociotechnical systems	Industrial automation	Mobility	Medicine	Financial services	Energy/environment
ISO TS 5083	Road vehicles – Safety for automated driving systems – Design, verification and validation	This document provides an overview and guidance of the steps for developing and validating an automated vehicle equipped with a safe automated driving system. The approach is based on top level safety goals and basic principles derived from worldwide applicable publications. It considers safety by design, verification and validation methods for automated driving focused on SAE level 3 and level 4 vehicles according to ISO/SAE PAS 22736. In addition, it outlines cybersecurity considerations throughout all described steps. The document is intended to be applied to road vehicles (incl. trucks and busses, i. e. road vehicles > 3,5to) excluding motorcycles.	ISO/TC 22/SC 32		X					X		
DIN EN ISO 22057	Sustainability in buildings and civil engineering works – Data templates for the use of environmental product declarations (EPDs) for construction products in building information modelling (BIM)	Formal integration of construction product data into BIM processes; AI reference as basis for data framework for ML/AI models.	NA 005-01-31-AA									X

Document	Title	Short description	Body	Relevance for key topics								
				Basic topics	Security/safety	Testing and certification	Sociotechnical systems	Industrial automation	Mobility	Medicine	Financial services	Energy/environment
DIN/TS 92004	Artificial intelligence – Quality requirements and processes – Risk scheme for AI systems throughout the entire life cycle	This document provides an AI risk scheme that covers risks along the entire life cycle of artificial intelligence (AI) systems that incorporate machine learning (ML) components. The scheme distinguishes between eight AI risk categories, i.e., reliability, fairness, autonomy and control, transparency, explainability, safety and security, and privacy, and assigns corresponding risk causes to each category. This document is intended to be applicable to all developers, providers, and operators of AI systems. It is intended to serve as a basis for identifying the AI risks that exist in a given AI system and for analyzing them, thus informing the parts of the risk management process within an organization that are responsible for identifying and analyzing risks.	NA 043-01-42-01 AK	X	X	X	X	X	X	X	X	X
ISO/IEC PWI 7699	Guidance for addressing security threats and failures in artificial intelligence		ISO/IEC JTC 1/SC 27 NA 043-04-27 AA	X	X	X	X	X	X	X	X	X

7.3 Bodies dealing with AI

Table 15 gives an overview of relevant standardization bodies dealing with AI.

Table 15: Overview of important AI standardization bodies[117]

	Body	Mirror body[118]
International	IEC/SyC AAL "System Committee AAL"	DKE/K 801
	IEC/TC 9 "Electrical equipment and systems for railways"	DKE/UK 351.3
	IEC/TC 44 "Safety of machinery – Electrotechnical aspects"	DKE/K 225
	IEC/SC 45A "Instrumentation, control and electrical power systems of nuclear facilities"	DKE/UK 967.1
	IEC/TC 62 "Medical equipment, software, and systems"	DKE/K 810
	IEC/TC 62/SC 62A "Common aspects of medical equipment, software, and systems"	DKE/UK 811.4
	IEC/TC 65/WG 10 "Security for industrial process measurement and control – Network and system security"	DKE/UK 931.1
	IEC/TC 65/SC 65A "System aspects"	DKE/GK 914
	ISO/CASCO "Committee on conformity assessment"	NA 147-00-03 AA
	ISO/IEC JTC 1/SC 7 "Software and systems engineering"	NA 043-01-07 AA
	ISO/IEC JTC 1/SC 27 "Information security, cybersecurity and privacy protection"	NA 043-04-27 AA
	ISO/IEC JTC 1/SC 32 "Data management and interchange"	NA 043-01-32 AA
	ISO/IEC JTC 1/SC 38 "Cloud computing and distributed platforms"	NA 043-01-38 AA
	ISO/IEC JTC 1/SC 41 "Internet of things and digital twin"	NA 043-01-41 AA
	ISO/IEC JTC 1/SC42 "Artificial Intelligence"	NA 043-01-42 GA
	ISO/TC 22/SC 32 "Electrical and electronic components and general system aspects"	NA 052-00-32 AA

[117] This overview makes no claim to completeness.

[118]
NA 005 DIN Standards Committee Building and Civil Engineering (NABau)
NA 023 DIN Standards Committee Ergonomics (NAErg)
NA 043 DIN Standards Committee on Information Technology and selected IT Applications (NIA)
NA 052 DIN Standards Committee Road Vehicle Engineering (NAAutomobil)
NA 053 DIN Standards Committee Rescue Services and Hospital (NARK)
NA 060 DIN Standards Committee Mechanical Engineering (NAM)
NA 063 DIN Standards Committee Medicine (NAMed)
NA 095 DIN Standards Committee Safety Design Principles (NASG)
NA 105 DIN Standards Committee Terminology (NAT)
NA 147 DIN Standards Committee Quality Management, Statistics and Certification (NQSZ)
NA 172 DIN Standards Committee Principles of Environmental Protection (NAGUS)
NA 175 DIN Standards Committee for Organizational Processes (NAOrg)

	Body	Mirror body[118]
	ISO/TC 22/SC 33 “Vehicle dynamics and chassis components“	NA 052-00-33 AA
	ISO/TC 23/SC 19 “Agricultural electronics“	NA 060-16-12 AA
	ISO/TC 37/SC 4 “Language resource management“	NA 105-00-06 AA
	ISO/TC 59/SC 17 “Sustainability in buildings and civil engineering works“	NA 005-01-31 AA
	ISO/TC 68 “Financial services“	NA 043-03-02 AA
	ISO/TC 121/SC1 “Breathing attachments and anaesthetic machines“	NA 053-03-01 AA
	ISO/TC 159/SC 1 “General ergonomics principles“	NA 023-00-01 GA
	ISO/TC 159/SC 3 “Anthropometry and biomechanics“	NA 023-00-03 GA
	ISO/TC 159/SC 4 “Ergonomics of human-system interaction“	NA 023-00-04 GA
	ISO/TC 163 “Thermal performance and energy use in the built environment“	NA 005-12-01 GA
	ISO/TC 176/SC 3 “Supporting technologies“	NA 147-00-01 AA
	ISO/TC 199 “Safety of machinery	NA 095 BR
	ISO/TC 204 “Intelligent transport systems“	NA 052-00-71 GA
	ISO/TC 207/SC 5 “Life cycle assessment“	NA 172-00-03 AA
	ISO/TC 210 “Quality management and corresponding general aspects for medical devices“	NA 063-01-13 AA
	ISO TC 215 “Health informatics“	NA 063-07-01 AA
	ISO/TC 262 “Risk management“	NA 175-00-04 AA
	ISO TC 276 “Biotechnology“	NA 063-09-02
	ISO/TC 299 “Robotics“	NA 060-38-01 AA
	ITU-T SG 13 “Future networks“	
	ITU-T SG 16 “Multimedia“	
European	CEN/CLC/JTC 13 “Cybersecurity and data protection“	NA 043-04-13 GA
	CEN/CLC/JTC 21 “Artificial Intelligence“	NA 043-01-42 GA
	CEN/TC 114 “Safety of machinery“	NA 095 BR
	CEN/TC 251 “Health Informatics“	NA 063-07-01 AA
	CLC/TC 62 “Electrical equipment in medical practice“	DKE/K 801
	ETSI “Methods for Testing and Specification (MTS)“	

	Body	Mirror body[118]
	ETSI “Securing Artificial Intelligence (SAI)“	
	ETSI “Speech and Multimedia Transmission Quality (STQ)“	
	ETSI “Experiential Networked Intelligence (ENI)“	
National	NA 159-07-01 AA “Financial services for the private household“	
	NA 175-00-03 AA “Social responsibility of organizations“	
	DKE/K 811 "General provisions for electrical equipment in medical use"	
	DKE/UK 931.1 "IT Security in automation technology”	
	DIN SPEC 2343 “Transmission of language-based data between artificial intelligences – Specification of parameters and formats“	
	DIN SPEC 13266 “Guideline for the development of deep learning image recognition systems“	
	DIN SPEC 92001 “Artificial Intelligence – Quality requirements and life cycle management for AI modules“	
	DIN SPEC 91426 “Quality requirements for video-based methods of personnel selection“	
	DIN SPEC 92001-3 “Artificial Intelligence – Life Cycle Processes and Quality Requirements – Part 3: Explainability”	
Consortia	IEEE AIMDWG “Artificial Intelligence Medical Device Working Group“	
	IEEE C/S2ESC – Software & Systems Engineering Standards Committee	
	IEEE C/SAB – Standards Activities Board	
	IEEE EMB/Stds Com – Standards Committee	
	IEEE RAS/SC – Standing Committee for Standards	
	IEEE SMC/SC – Standards Committee	
	IEEE SSIT/SC – Social Implications of Technology Standards Committee	
	IEEE VT/ITS “Intelligent Transportation Systems“	

8
Index of abbreviations

Abbreviation	Meaning
AAS	Asset Administration Shell
ADM	Algorithmic Decision Making
AG	Arbeitsgruppe (Working Group)
AGV	Automated Guided Vehicles
AI	Artificial Intelligence
AI Act	Artificial Intelligence Act
AIM	AI Machine
AIMS	AI Management System
ALKS	Automated Lane Keeping System
API	Application Programming Interface
AR	Augmented Reality
ArbMedVV	Arbeitsmedizinische Vorsorge Verordnung (Occuplational health precaution ordinance)
ArbStättV	Arbeitsstättenverordnung (German Workplace Ordinance)
ASR	Automatic Speech Recognition
ATDD	Acceptance-Test-Driven Development
AUC	Area under the Curve
AUGT	Automatischer städtischer schienengebundener Personennahverkehr (Automated urban rail-based public transport system)
AV	Aerial Vehicles
AVP	Valet Parking System
B2B	Business to Business
BaFin	Bundesanstalt für Finanzdienstleistungsaufsicht (German Federal Financial Supervisory Authority)
BAIT	Bankaufsichtliche Anforderungen an die IT (Banking supervisory IT requirements)

Abbreviation	Meaning
BetrSichV	Betriebssicherheitsverordnung (Industrial Safety Ordinance)
BIM	Building Information Modelling
BMAS	Bundesministerium für Arbeit und Soziales (Federal Ministry of Labour and Social Affairs)
BMBF	Bundesministerium für Bildung und Forschung (Federal Ministry of Education and Research)
BMUV	Bundesministerium für Umwelt, Naturschutz, nukleare Sicherheit und Verbraucherschutz (Federal Ministry for the Environment, Nature Conservation, Nuclear Safety and Consumer Protection)
BMVI	Bundesministerium für Verkehr und digitale Infrastruktur (Federal Ministry of Transport and Digital Infrastructure)
BMWK	Bundesministerium für Wirtschaft und Klimaschutz (Federal Ministry for Economic Affairs and Climate Action)
BSI	Bundesamt für Sicherheit in der Informationstechnik (Federal Office for Information Security)
CC	Common Criteria
CCAM	Cooperative, Connected und Automated Mobility
CC-KING	Competence Center KI-Engineering (Competence Centre AI Systems Engineering)
CCRA	Common Criteria Recognition Arrangement
CNN	Convolutional Neural Networks
COLREG	Convention on the international regulations for preventing collisions at sea
CPU	Central processing units
CRR	Capital Requirements Regulation

Abbreviation	Meaning
CSM-RA	Common Safety Method for Risk Evaluation and Assessment
D&A	Detect and Avoid
DER	Distributed Energy Resources
DGUV	Deutsche Gesetzliche Unfallversicherung (German Social Accident Insurance)
DICOM	Digital Imaging and Communications in Medicine
DKE	Deutsche Kommission Elektrotechnik Elektronik Informationstechnik in DIN und VDE (German Commission for Electrical, Electronic & Information Technologies of DIN and VDE)
DL	Deep Learning
DPP	Digital Product Passport
DSGVO	Datenschutz-Grundverordnung (GDPR – General Data Protection Regulation)
DSO	Distribution System Operator (Verteilnetzbetreiber)
EAD	Ethically Aligned Design
EAL	Evaluation Assurance Levels
EBA	European Banking Authority
EHDS	European Health Data Space
EHF	Ergonomics/Human Factors
ELGI	Ethische Leitlinien der Gesellschaft für Informatik e. V. (Ethical guidelines of the Society for Informatics)
ENISA	European Union Agency for Cybersecurity
EOSC	European Open Science Cloud
ESG	Environmental Social Governance
EU	European Union
EV	Electric Vehicle

Abbreviation	Meaning
F&E	Forschungs- und Entwicklungsarbeiten (R&D = Research and development)
FAIR	Findable, Accessible, Interoperable, Reusable
FDA	Food and Drug Administration
GAN	Generative Adversial Networks
GCP	Good Clinical Practice
GefStoffV	Gefahrstoffverordnung (Hazardous Substances Ordinance)
GIS	Geographic Information System
GoA	Grade of Automation
GPU	Graphics processing units
HAS	Harmonized Standards
hEN	Harmonized European Standard
HIC	Human in Command
HITL	Human-in-the-Loop
HLEG	High Level Expert Group
HMI	Human-machine interface
HOTL	Human-on-the-Loop
HR	Human Resources
IACS	Industrial Automation and Control Systems
IEEE	Institute of Electrical and Electronics Engineers
IETF	Internet Engineering Task Force
IG-NB	Interessensgemeinschaft der Benannten Stellen (Association of Notified Bodies)
IKT	Informations- und Kommunikations-technologien (ICT = Information and Communications Technologies)
IML4E	Industrial Grade Machine Learning for Enterprises

Abbreviation	Meaning
IMO	International Maritime Organization
IoU	Intersection over Union
IRB(A)	Internal ratings-based (approach)
ISMS	Information Security Management System
IVD	In-vitro diagnostics
IVDR	In-vitro Diagnostic Medical Devices Regulation
KAIT	Kapitalverwaltungsaufsichtliche Anforderungen an die IT (Capital management supervisory requirements for IT)
KAMaRisk	Mindestanforderungen an das Risikomanagement von Kapitalverwaltungsgesellschaften (Minimum requirements for the risk management of capital management companies)
KI	Künstliche Intelligenz (AI = artificial intelligence)
KPIs	Key Performance Indicators
KRITIS	Kritische Infrastrukturen (Critical infrastructures)
KTIs	Key Trustworthiness Indicators
LCA	Life Cycle Assessment
LoD	Level of Detail
LROD	Long Range Obstacle Detection
LSA	Lichtsignalanlagensteuerung (Traffic signal control)
MaRisk	Mindestanforderungen für das Risikomanagement für deutsche Kreditinstitute (Minimum requirements for risk management for German credit institutions)
MDR	Medical Device Regulation
ML	Machine Learning
MRT	Magnetic Resonance Tomography

Abbreviation	Meaning
MSS	Management System Standard
MTO	Mensch, Technik und Organisation (Human, technology and organization)
NFDI	Nationale Forschungsdateninfrastruktur (National Research Data Infrastructure)
NRM KI	Normungsroadmap Künstliche Intelligenz (Standardization Roadmap Artificial Intelligence)
ODD	Operational Design Domain
OEM	Original Equipment Manufacturer
OGC	Open Geospatial Consortium
OMG	Object Management Group
OWL	Web Ontology Language
PACS	Picture Archiving and Communication System
PMS	Power Management System
POC	Probability of Classification
POD	Probability of Detection
QML	Quantum Machine Learning
RAM	Reference Architecture Model
RAMI 4.0	Reference Architecture Model Industrie 4.0
RDF	Resource Description Framework
SAE	Society of Automotive Engineers
SG	Smart Grid
SGAM	Smart Grid Architecture Model
SIF	System Interface
SM	Smart Manufacturing
SOTIF	Safety of the intended Function
TAI	Trusworthy Artificial Intellligence

Abbreviation	Meaning
Tf	Triebfahrzeugführenden (Train driver)
TRM	Trustworthiness Readiness Matrix
UAM	Urban Air Mobility
UML	Unified Modeling Language
VAIT	Versicherungsaufsichtliche Anforderungen an die IT (Insurance supervisory requirements for IT)
VNB	Verteilnetzbetreiber (Distribution network operator)
VWS	Verwaltungsschale (Administration shell)
W3C	World Wide Web Consortium
XAI	Explainable AI
ZAIT	Zahlungsdiensteaufsichtliche Anforderungen an die IT (Payment services supervisory requirements for IT)
ZFP	Zerstörungsfreie Prüfung (NDT = Non-destructive testing)

9 Glossary

Term	Meaning and use
accessibility	Extent to which products, systems, services, environments, and facilities can be used by people from a population with the broadest range of user needs, characteristics, and skills to achieve identified goals in identified contexts of use. Note on terminology: The context of use includes direct use or use supported by assistance technologies. Note on German translation: The terms "Barrierefreiheit" ("barrier-free") and "Zugänglichkeit" ("accessible") are often used interchangeably. "Barrierefreiheit" is more than just physical accessibility, but includes that. This is why, for example, in German the term "Zugänglichkeit" is preferred in the construction sector and the term "Barrierefreiheit" in the ICT sector. Mentioned in DIN EN ISO 9241-210:2020 [183], ISO 9241-112:2017 [249].
accountability	Describes a relationship between an actor and a forum in which the actor has to explain and justify their position. The forum has the right to question the operator's explanations (for clarification and additional explanations) and to make a judgement. Basically, consequences should be announced to the actor so that accountability is perceived and implemented by the actor. Mentioned in ISO/IEC 22989:2022 [16].
accreditation	Confirmation by a third party formally stating that a conformity assessment body has the competence, impartiality as well as uniform operation to perform certain conformity assessment activities.
accreditation body	Authorized body that performs accreditations.
accuracy (in the context of classification)	In the context of classification in AI, accuracy is a metric for measuring the quality of mostly binary classifications. It is calculated as the proportion of correct classifications to all classifications. Mentioned in ISO/IEC TR 29119-11 [132], ISTQB – CTAI Syllabus v1.0 [137].
adaptability	Ability of a system to respond to changes in its environment to continue to fulfil both functional and non-functional requirements. Mentioned in ISO/IEC TR 29119-11 [132].
adversarial attack	An adversarial attack is a deliberate attempt to cause errors using adversarial examples. Artificial neural networks in particular are considered especially vulnerable to this type of attack. Mentioned in ISO/IEC TR 29119-11 [132], ISTQB – CTAI Syllabus v1.0 [137].
agency	Agency is an "enactment" of political-ethical forms of subjectivity. The focus is on the processuality of iterative practices ("doing"). Agency is not only reserved for humans, but can also be attributed to non-human entities.
agent	In the context of AI, an agent is understood to be a decisive and acting system that can interact with its environment and other agents Mentioned in ETSI GR ENI 004 V2.2.1 [499], ISO/IEC 22989:2022 [16].
AI capability	Ability such as "perceiving", "acting" or "communicating", which is implemented on the basis of artificial intelligence methods. See Chapter 4.1.1.1.
AI component	A component that incorporates AI methods. Mentioned in ISTQB – CTAI Syllabus v1.0 [137].

Term	Meaning and use
AI module	Software module in which AI methods are implemented. AI services as building blocks in a chain of supply relationships involving multiple IT components or AI services (see Chapter 4.3.2.1).
AI system	System that uses artificial Intelligence.
application programming interface (API)	A set of communication protocols, code, and tools that enables a set of software components to interact with either a human or another set of software components. Mentioned in ETSI GR ENI 004 V2.2.1 [499].
approval	Permission to market or use a product, service or process for the stated purpose or under stated conditions.
artificial intelligence (AI)	The term is discussed in different disciplines from different perspectives. Due to the AI effect, the term is constantly evolving. Three definitions are shown below: Definition 1: Ability of a technical system to acquire, process and apply knowledge and competencies (ISO/IEC TR 29119-11) [132]. Definition 2: A computer-based system that operates cognitively to understand information and solve problems (ISO/IEC 22989:2022 [16]). Definition 3: Artificial intelligence refers to a family of technologies, [...] and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with (European draft AI Act, [4]). Mentioned in ETSI GR ENI 004 V2.2.1 [499], ISO/IEC 22989:2022 [16], ISO/IEC TR 29119-11 [132], ISTQB – CTAI Syllabus v1.0 [137] 0.
artificial neural network (ANN)	ANNs are networks of artificial neurons and have a biological model. Borrowing from biology, an artificial neuron is an object that responds to one or more stimuli, depending on how strongly it is activated or the stimulus is weighted. An ANN basically consists of an input layer and an output layer. In between are hidden layers or activity layers. ANNs usually always need to be trained before they can solve problems. In this process, a particular algorithm or the neural network weights the connections of the neurons based on given learning material and learning rules until it has reached or developed a certain learning goal. Mentioned in ETSI GR ENI 004 V2.2.1 [499].
audit	Process for obtaining relevant information about an object of conformity assessment and evaluating it objectively in order to determine the extent to which the specified requirements are met Note 1 to entry: Examples of objects of an audit are management systems [...]. Note 2 to entry: Only organizations are audited, not products or services.
autonomous system	A system that works for extended periods of time without human intervention. Mentioned in ISO/IEC TR 29119-11 [132], ISTQB – CTAI Syllabus v1.0 [137].
autonomy	Autonomy is the absence of heteronomy. In relation to humans, autonomy means free will and corresponds to a basic principle of digital ethics. Mentioned in ISO/IEC 22989:2022 [16], ISO/IEC TR 29119-11 [132], ISTQB – CTAI Syllabus v1.0 [137].

Term	Meaning and use
availability	The property of being accessible and usable by authorized persons when needed. Characterized by degree, the extent of availability may depend on features such as timeliness, interpretability, as well as completeness of information. Mentioned in ISO/IEC 22989:2022 [16].
Bayesian network	A Bayesian network is a directed, cycle-free graph. While in the graph the nodes represent variables with ranges of values, the edges represent conditional probabilities. Mentioned in ISO/IEC 22989:2022 [16].
bias	General: The deviation from a reference value or the actual value. In the context of AI, bias is often understood as a systematic deviation that does not correspond to the actual or desired distribution. In AI applications, an existing bias is often seen as being unfair to a particular person or group. A bias can have its cause in data, an algorithm itself, sociocultural influences, or any combination of the aforementioned causes. Against this background, human cognitive distortions are also part of the bias concept. Mentioned in ETSI GR ENI 004 V2.2.1 [499], ISO/IEC 22989:2022 [16], ISO/IEC TR 29119-11 [132], ISTQB – CTAI Syllabus v1.0 [137].
big data	Data whose characteristics in terms of volume, complexity, change dynamics and/or lack of structure require special technologies, techniques and methods for processing. Mentioned in ISTQB – CTAI Syllabus v1.0 [137].
building information modelling (BIM)	Working method for the networked design and construction of buildings using information-based models.
certification	Confirmation by a third party, related to an object of conformity assessment, except accreditation.
chatbot	An application from computational linguistics for conducting a text-based conversation on text or synthesis of natural language. Mentioned in ISTQB – CTAI Syllabus v1.0 [137].
classification (machine learning)	Task by means of which the output class for a given input is predicted. Mentioned in ISO/IEC TR 29119-11 [132], ISTQB – CTAI Syllabus v1.0 [137].
classifier	A method /system used to implement a classification task. Mentioned in ETSI GR ENI 004 V2.2.1 [499], ISO/IEC TR 29119-11 [132], ISTQB – CTAI Syllabus v1.0 [137].
closed-box testing	A closed-box test (also black-box test) is a test procedure in which no internals (such as the AI model in particular) of the AI system are available to a tester for test scenarios. In contrast, usually only inputs to the AI system with associated outputs from the AI system are available. Mentioned in ISO/IEC TR 29119-11 [132].
cognition	Understanding data and information and generating new data, information and new knowledge. Mentioned in ETSI GR ENI 004 V2.2.1 [499].
completeness	Degree to which data associated with an entity has values for all attributes of that entity and for entities related to it.
computer linguistics	Computer linguistics studies how natural language in the form of text or speech data can be processed algorithmically with the help of computers. It is the interface between linguistics and computer science.

Term	Meaning and use
computer vision	AI capability of a functional unit to acquire, process and interpret visual data. Computer vision involves the use of sensors to create a digital image of a visual scene. See Chapter 4.1.2.6. Mentioned in ISO/IEC 22989:2022 [16].
confirmation	Creating a statement based on a decision that compliance with specified requirements has been demonstrated.
conformity assessment	Demonstration that specified requirements are met.
conformity assessment body	Body performing conformity assessment activities, but not accreditation.
continual learning	In the context of AI, continual learning is the training of an AI system that occurs iteratively and incrementally in parallel with its operation. Mentioned in ISO/IEC 22989:2022 [15], [16].
control (in the context of certification as well as security, safety or data protection)	Procedures (technical, organizational, legal, physical) to mitigate risks for security, safety, or privacy issues.
controllability	Property by means of which a human or other external agent can directly and immediately intervene in the ongoing function of the system. Mentioned in ISO/IEC 22989:2022 [16].
criticality	Measure of the potential hazards that can arise from the use of an AI system in a specific application context. The term is often used in a similar way to risk, with criticality being more focused on an assessment of the system as a whole.
currentness	Degree of temporal validity of data with relevance for a specific application context.
data mining	Computer-based process in which patterns are extracted from various dimensions by means of the analysis of quantitative data, and are categorized, and potential relationships and consequences are identified. Mentioned in ETSI GR ENI 004 V2.2.1 [499], ISO/IEC 22989:2022 [16].
data poisoning	The intentional and malicious manipulation of training, validation, testing, or input data for AI models. Mentioned in ISTQB – CTAI Syllabus v1.0 [137].
data quality	Degree to which characteristics of data meet explicitly specified or implicit requirements for a given use case.
dataset	Collection of data with a common format and target-relevant content. Ideally, the data selected in this way represents the larger dataset or assumed real-world characteristic. Note: Datasets can be used for training, validation and testing of an AI model. In the context of supervised machine learning, datasets provide a basis for training the learning algorithm. Example 1: Microblogging posts from June 2020 linked to the hashtags #rugby and #football. Example 2: Macro photos of flowers with size 256x256 pixels. Mentioned in ISO/IEC 22989:2022 [16], ISTQB – CTAI Syllabus v1.0 [137].
declaration (testing and certification)	Confirmation by a first party (for example, self-declaration of a manufacturer).

Term	Meaning and use
deep learning (DL)	Deep Learning refers to a class of artificial neural network optimization methods (see Artificial Neural Network) that have numerous hidden layers between the input layer and the output layer and thus have an extensive internal structure. As an extension of learning algorithms for network structures with very few or no intermediate layers, deep learning methods enable stable learning success even with numerous intermediate layers. Mentioned in ETSI GR ENI 004 V2.2.1 [499], ISO/IEC 22989:2022 [16], ISO/IEC TR 29119-11 [132], ISTQB – CTAI Syllabus v1.0 [137].
deep neural network	Neural network that, in addition to the input and output layers, has other, “hidden” layers of nodes (cf. “deep learning”). Mentioned in ISO/IEC TR 29119-11 [132], ISTQB – CTAI Syllabus v1.0 [137].
digital twin	Virtual digital representation of a physical object or system over its life cycle using real-time data. The digital representation can be included as a basis for traceability, training, and inference of AI models. Mentioned in ETSI GR ENI 004 V2.2.1 [499].
ergonomics	Scientific discipline concerned with understanding the interactions between human elements and other elements of a system. Furthermore, also a profession that applies theory, principles, data, and methods to the design of work systems with the goal of optimizing human well-being and overall system performance. Note: This definition is consistent with that established by the International Ergonomics Association. Mentioned in DIN EN ISO 26800:2011 [239].
ethics	Principles that determine the moral behaviour of a human being or a machine (according to ETSI). Across domains, ethics is the scientific study of morality. It reflects and philosophizes on diverse moral concepts, it analyzes and systematizes them, it examines and questions their justifications and principles. There are various moral concepts, systems of norms, principles, values or dispositions, all of which claim to be the basis for right action. Mentioned in ETSI GR ENI 004 V2.2.1 [499].
examinability	The ability to track statements, for example by granting access to data, documents or (AI) systems
expert system	Often rule-based system based on symbolic knowledge processing. Example: If-then rules. Note: E.g., symbolic, formal representation of knowledge in AI systems with the property of inferring new knowledge from formal knowledge by means of reasoning based on logic. Mentioned in ISO/IEC 22989:2022 [16], ISTQB – CTAI Syllabus v1.0 [137].
explainability	Desirable property of an AI system in that factors that led to an automated decision by the system can be “understood” by a human. Mentioned in ISO/IEC 22989:2022 [16], ISO/IEC TR 29119-11 [132], ISTQB – CTAI Syllabus v1.0 [137].
explainable AI (XAI)	A research and application area concerned with understanding the factors that influence outcomes of AI systems. Mentioned in ISTQB – CTAI Syllabus v1.0 [137].

Term	Meaning and use
fairness	In the use of algorithmic and sociotechnical systems in the broader sense and machine-learning systems in the narrower sense, fairness as an ethical principle describes the reproducible degree of equal treatment of different people in all stages of the system's life cycle. This principle is also applicable to non-human actors (e.g. animals, environment, nature) or to natural actors in general.
false negative (FN)	A model prediction where the model of a binary classification incorrectly predicts negative when positive would be correct. Mentioned in ISO/IEC TR 29119-11 [132], ISTQB – CTAI Syllabus v1.0 [137].
false positive (FP)	A model prediction where the model of a binary classification incorrectly predicts positive when negative would be correct. Mentioned in ISO/IEC TR 29119-11 [132], ISTQB – CTAI Syllabus v1.0 [137].
feature	Individually measurable property of an object under observation. Mentioned in ETSI GR ENI 004 V2.2.1 [499], ISTQB – CTAI Syllabus v1.0 [137].
general AI	AI that exhibits intelligent behaviour comparable to that of a human across the spectrum of AI cognitive capabilities (synonym: strong AI). Mentioned in ISO/IEC 22989:2022 [16], ISO/IEC TR 29119-11 [132], ISTQB – CTAI Syllabus v1.0 [137].
glass box testing	A glass box test (also white box test) is a test procedure in which a tester has internals (such as the AI model) of the AI system available to generate the test cases. Mentioned in ISO/IEC TR 29119-11 [132].
graph	A mathematical model that represents connecting structures in an abstract way. It consists of "nodes" and "edges" that represent connections between these nodes. Both nodes and edges can be assigned values depending on the application, for example weights, costs or distances. Mentioned in ETSI GR ENI 004 V2.2.1 [499].
graphical processing unit (GPU)	An application-specific integrated circuit with optimized memory utilization to accelerate the generation of images in an image buffer. Mentioned in ISO/IEC TR 29119-11 [132], ISTQB – CTAI Syllabus v1.0 [137].
ground truth	Information obtained by direct observation and measurement and assumed to be real or true. Mentioned in ISO/IEC 22989:2022 [16], ISTQB – CTAI Syllabus v1.0 [137].
human-centred design	Approach to system design and development that aims to make interactive systems more usable by focusing on the use of the system and applying knowledge and techniques from the fields of occupational science/ergonomics and usability. Mentioned in DIN EN ISO 6385:2016 [235], ISO 9241-210:2020 [183].

Term	Meaning and use
hyperparameter	In machine learning, hyperparameters usually refer to all parameters that are not directly defined or influenced by the training process. This includes model parameters such as the number of layers of a neural network or the step size of the training process, but not, for example, the weights learned. Basically, hyperparameters can be differentiated between algorithmic and model-specific hyperparameters. Algorithmic hyperparameters affect the performance of the learning algorithm; whereas model-specific hyperparameters affect the mathematical or statistical model used in the learning process. Mentioned in ETSI GR ENI 004 V2.2.1 [499], ISO/IEC 22989:2022 [16], ISO/IEC TR 29119-11 [132], ISTQB – CTAI Syllabus v1.0 [137].
information security	→ See security (within the meaning of IT security)
inspection	Examination of an object of conformity assessment and determination of its conformity with detailed requirements or, on the basis of expert assessment, with general requirements.
internet of things (IoT)	The Internet of Things (IoT) networks a variety of diverse (edge) devices (see also IoT device) and central data platforms, thus connecting systems, services, people and information from the physical and virtual worlds. In addition to new applications and services, the IoT has also enabled the development of new business models. Mentioned in ISO/IEC 22989:2022 [16].
interpretability	The degree of comprehensibility of the functioning of an underlying (AI) technology. Mentioned in ISO/IEC TR 29119-11 [132], ISTQB – CTAI Syllabus v1.0 [137].
knowledge representation	Representation of knowledge that is usable by an AI system, e.g., an expert system. Mentioned in ETSI GR ENI 004 V2.2.1 [499].
label/annotation	In machine learning, labels or annotations are the parts of the training dataset that specify the desired ideal output of the model for a corresponding input for training purposes. In a broader sense, this also refers to the actual output of a model in operation. Mentioned in ETSI GR ENI 004 V2.2.1 [499], ISO/IEC 22989:2022 [16].
learning algorithm	An algorithm that builds an ML model based on characteristics of the training datasets. → See learning system Mentioned in ISTQB – CTAI Syllabus v1.0 [137].
learning data	→ See training data
learning system	Learning systems are machines, robots, and software systems that autonomously perform abstractly described tasks based on data that serve as their learning basis, without each step being specifically programmed by humans. To solve tasks, they use models trained by learning algorithms. With the help of the learning algorithm, many systems can continue to learn during operation (continual learning): They improve the models they trained in advance and expand their knowledge base. → See learning algorithm.
life cycle	Time course used to characterize a system, product, service, project, or other human-made entity from conception to decommissioning. Mentioned in ISO/IEC 22989:2022 [16]

Term	Meaning and use
life cycle assessment (LCA)	Determination of the inventory and associated environmental impacts of a product/service.
machine learning (ML)	ML, as a subfield of AI and an umbrella term for the "artificial" generation of knowledge, employs computational techniques to enable systems to learn from data or experience. Such a system can generalize the acquired knowledge after the end of the learning phase by recognizing patterns and regularities from the learning data and transferring them to unknown data (learning transfer). Mentioned in ETSI GR ENI 004 V2.2.1 [499], ISO/IEC 22989:2022 [16], ISO/IEC TR 29119-11 [132], ISTQB – CTAI Syllabus v1.0 [137].
machine translation	Automatic translation of spoken or written natural language into another language by an AI system. Mentioned in ISO/IEC 22989:2022 [16].
metric	A metric is a measure to quantify the property of an object. In the field of AI or machine learning, it is used to measure the characteristics of an AI system and thus map them into key figures that are as informative as possible The key figures can refer to quality criteria such as a false positive rate for classification outputs or the mean square error for regression tasks. In addition, they can present more advanced evaluation criteria such as the strength of the bias between genders. The metric should be defined via an algorithmically implementable measurement principle so that it can be applied to concrete problems. Note: Metrics that evaluate how well an AI system performs its task or function are also called functional performance metrics. Mentioned in ETSI GR ENI 004 V2.2.1 [499], ISO/IEC TR 29119-11 [132].
ML model	A mathematical construct that makes a conclusion or prediction based on input data. Mentioned in ETSI GR ENI 004 V2.2.1 [499], ISO/IEC 22989:2022 [16].
ML system	A system that integrates ML models. Mentioned in ISTQB – CTAI Syllabus v1.0 [137].
model	Physical, mathematical, or otherwise logical representation of a system, entity, phenomenon, process, or data, including their relationships and dependencies, using a specified set of rules and concepts. Mentioned in ETSI GR ENI 004 V2.2.1 [499], ISO/IEC 22989:2022 [16], ISO/IEC TR 29119-11 [132].
ontology	On the one hand, ontology is a philosophical discipline that deals with the classification of concepts of the world into category systems that are as meaningful as possible. On the other hand, in computer science such category systems, for example consisting of terms and relations for algorithmic use, are concretely called "ontologies". Mentioned in ETSI GR ENI 004 V2.2.1 [499].
overfitting	Overfitting is when an ML model is so heavily biased towards the training data set that it is difficult to generalize to new data. Mentioned in ISO/IEC TR 29119-11 [132], ISTQB – CTAI Syllabus v1.0 [137].
parameter	In the context of machine learning: internal variable of a model that affects the computation of results. Mentioned in ISO/IEC 22989:2022 [16], ISO/IEC TR 29119-11 [132].
path optimization	Algorithm for identifying the shortest/most favourable path in a graph of nodes and edges.

Term	Meaning and use
planning	AI method that assembles a workflow from a series of actions to achieve a specific goal. Mentioned in ISO/IEC 22989:2022 [16].
precision (in the context of classification)	Precision or "positive predictive value" in the context of classification is the proportion of a system's correct-positive outputs to its total positive outputs. Metnioned in ISO/IEC TR 29119-11 [132], ISTQB – CTAI Syllabus v1.0 [137].
predictability	Property of an AI system that enables reliable assumptions about results or the quality of predictability. The degree can be used to describe the extent of accurate speculation about conditions and processes that have occurred. Mentioned in ISO/IEC 22989:2022 [16].
prediction	Function of an ML model that leads to a predicted target value for a given input. Mentioned in ISO/IEC 22989:2022 [16], ISO/IEC TR 29119-11 [132].
recall	In formal contexts, recall or "true-positive rate" refers to the proportion of true-positive outputs of a system among the actual positive target outputs (see true positive as well as false negative). Mentioned in ISO/IEC TR 29119-11 [132], ISTQB – CTAI Syllabus v1.0 [137].
refurbishment track	Optimized sequence of renovations in the construction sector.
regression	ML method that results in a quantitative output value for a given input. Mentioned in ISO/IEC TR 29119-11 [132], ISTQB – CTAI Syllabus v1.0 [137], ISO/IEC 23053:2022 [24].
reinforcement learning	Using software agents to perform actions in an environment with the goal of maximizing a cumulative reward. Mentioned in ETSI GR ENI 004 V2.2.1 [499], ISO/IEC 22989:2022 [16], ISO/IEC TR 29119-11 [132], ISTQB – CTAI Syllabus v1.0 [137].
reliability	The property of being able to exhibit trustworthy behaviour and the property of consistently exhibiting intended behaviour and results. Mentioned in ISO/IEC 22989:2022 [16].
resilience	Resistance to disturbances and failures with the associated ability to prevent parasitic influences on the operation. Mentioned in ISO/IEC 22989:2022 [16].
retraining	Updating a trained model by training it again with different training data. Mentioned in ISO/IEC 22989:2022 [16].
risk	The term risk usually refers to undesirable events with as yet uncertain occurrence that are associated with a product or process; in formal definitions, they are usually characterized as a combination of the amount of damage and the probability of occurrence of a loss. The overall quantitative classification is usually derived from the expected value for the damage (to be specified in more detail in each case), i.e. as the product of the amount of damage and the probability of occurrence. Sometimes additional parameters are integrated (e.g. probability of detection or avoidability by human intervention). Mentioned in ISO/IEC 22989:2022 [16].
robot	A robot is a technical system that has sensors to perceive its environment, a purpose-oriented processing unit, and effectors to change its spatial relation in the environment or to change the environment itself. Mentioned in ISO/IEC 22989:2022 [16], ISO/IEC TR 29119-11 [132].
robotics	Discipline that deals with the design and construction of robots. Mentioned in ISO/IEC 22989:2022 [16].

Term	Meaning and use
robustness	Ability of a system to perform its function under any circumstances. Mentioned in ISO/IEC 22989:2022 [16].
safety	Safety usually refers specifically to the absence of risks to life and limb posed by a system. In a broader sense, mental health and the integrity of the environment and other values are also counted as safety. Mentioned in ISO/IEC 22989:2022 [16], ISO/IEC TR 29119-11 [132], ISTQB – CTAI Syllabus v1.0 [137].
security (within the meaning of information security)	The term information security (the term "security" alone inadequately reflects this) refers to the ability of a system, over its life cycle, to, among other things, protect important information from unauthorized access, ensure its availability, or preserve its confidentiality, integrity, authenticity, accountability, and reliability, including for functionality. Mentioned in DIN EN ISO/IEC 27000 series [131].
semantics	Research field about analyzing the meaning of something (e.g., a sentence or a relation in a model). Mentioned in ETSI GR ENI 004 V2.2.1 [499].
semi supervised ML	Hybrid of supervised and unsupervised learning, where the training data consists of both labelled and unlabelled data. Mentioned in ETSI GR ENI 004 V2.2.1 [499], ISO/IEC 22989:2022 [16].
smart grid	Combination of energy technology with information and communications technology to optimize energy generation, transport and use.
smart grid architecture model	Generic model of a smart grid, with the help of which the implementation possibilities of various services or functions are examined.
sociotechnical system	Sociotechnical systems include the subsystems of humans and technology, which are interrelated and interact or should interact with each other. In this context, AI technology is related to humans, the organizational environment, and society as a whole.
specified requirement	Demand or expectation that is set down.
speech recognition	An AI capability that uses the conversion of a speech signal from voice to text to represent the content of the speech. Mentioned in ISO/IEC 22989:2022 [16].
stress (person-related)	A person's internal response to stress (loads), depending on their individual characteristics (e.g., height, age, abilities, talents, skills, etc.). Note 1: In DIN EN ISO 6385:2016 [235] "stress" is expressed as "work strain". Note 2: The term "stress" is neutral. Its effects can be positive, neutral or negative. Mentioned in DIN EN ISO 26800:2011 [239].
structured data	Information organized in a particular way (a fixed format, data model, or schema) in a dataset or file. Mentioned in ETSI GR ENI 004 V2.2.1 [499].
subsymbolic AI	Type of AI methods based on models with numerical representation and implicit information coding. Mentioned in ISO/IEC 22989:2022 [16].

Term	Meaning and use
supervised ML	In a narrower sense, supervised learning refers to machine learning methods that are trained with concretely specified target outputs (“labels”). In a broader sense, this includes methods whose learning target is determined by concrete provisions, even if not at the level of individual issues. This broader sense includes practices such as GANs and reinforcement learning. Mentioned in ISO/IEC 22989:2022 [16].
support vector machine (SVM)	Machine learning method that finds decision boundaries with maximum limit value. Mentioned in ISO/IEC 22989:2022 [16], ISTQB – CTAI Syllabus v1.0 [137].
symbolic AI	A type of AI method based on processing symbols and structures. Mentioned in ISO/IEC 22989:2022 [16].
syntax	Set of rules that determine how elements of a statement are structured. Mentioned in ETSI GR ENI 004 V2.2.1 [499].
target population	Group of people for whom something is designed, described in terms of relevant features. Mentioned in DIN EN ISO 26800:2011 [239].
taxonomy	Method of classifying objects according to certain criteria.
test data (in the context of AI)	Data used to evaluate the performance of a final AI model (in general) or machine learning model (in particular) before it goes live. Note: In principle, test data should be separated from training data and validation data. Mentioned in ISO/IEC 22989:2022 [16], ISO/IEC TR 29119-11 [132].
testing (in the meaning of testing and certification)	Determination of one or more features on an object of conformity assessment according to a procedure.
training	Process of imparting a set of knowledge, AI skills, procedures, and/or behaviours to an entity. Mentioned in ETSI GR ENI 004 V2.2.1 [499], ISO/IEC 22989:2022 [16].
training data	Data that can be used in the training process to create an AI model. Mentioned in ISO/IEC 22989:2022 [16], ISO/IEC TR 29119-11 [132].
transformer	In machine learning, transformers and their architectures belong to neural models that are used for numerous language technology tasks, among others. Transformers belong to the deep learning architectures.
transparency	Availability of an open, understandable, and accessible representation of information about functional aspects of an AI system. This includes, among other things, the explainability of the AI system (e.g., neural networks), the comprehensibility of the data protection concept, and information on quality assurance processes during development. Mentioned in ISO/IEC 22989:2022 [16], ISO/IEC TR 29119-11 [132], ISTQB – CTAI Syllabus v1.0 [137].
true negative (TN)	A prediction where the model correctly predicts the negative category. Mentioned in ISO/IEC TR 29119-11 [132], ISTQB – CTAI Syllabus v1.0 [137].

Term	Meaning and use
true positive (TP)	A prediction where the model correctly predicts the positive category. Mentioned in ISO/IEC TR 29119-11 [132], ISTQB – CTAI Syllabus v1.0 [137].
trustworthiness	The ability to demonstrably meet expectations. Note 1: Depending on the context or sector, and also on the specific product or service, the data, and the technology used, features that need to be reviewed to ensure that stakeholder expectations are met differ. Note 2: Trustworthiness characteristics include, for example, reliability, availability, resilience, security and safety, privacy, accountability, transparency, integrity, authenticity, quality, and usability. Note 3: Trustworthiness is an attribute that can be applied to services, products, technology, data and information, and – in the context of governance – to organizations. Mentioned in ISO/IEC 22989:2022 [16].
underfitting	The creation of an ML model that does not reflect the underlying trend of the training dataset, resulting in a model that has difficulty making accurate predictions. Mentioned in ISO/IEC TR 29119-11 [132], ISTQB – CTAI Syllabus v1.0 [137].
understandability	The property of an entity, system, or process to be traceable.
unsupervised ML	Unsupervised learning refers to machine learning methods that learn a function without relying on concretely specified targets (for example, “labels”). There are different opinions as to the degree of concreteness of external targets that can no longer be considered “unsupervised learning”. Mentioned in ETSI GR ENI 004 V2.2.1 [499], ISO/IEC 22989:2022 [16], ISO/IEC TR 29119-11 [132], ISTQB – CTAI Syllabus v1.0 [137].
usability	Extent to which a system, product, or service can be used by certain users in a certain context of use to achieve certain goals effectively, efficiently, and satisfactorily. Note 1: The “certain” users, “certain” goals, and “certain” context of use refer to the respective combination of users, goals, and context of use that are assumed to be usable. Note 2: The word “usability” is also used as a qualifier to refer to design knowledge, skills, activities, and attributes that contribute to usability, such as usability expertise and usability professionals, usability-oriented development, processes and evaluation, and usability heuristics. Mentioned in DIN EN ISO 9241-210:2020 [183].
validation (in the context of ML)	Validation, also referred to as ML model optimization or model tuning, in the context of ML refers to the testing of trained ML models using validation data. This allows the quality of the trained ML models to be identified, compared and optimized (→ see hyperparameters). In particular, it is usually possible to see whether the ML model can generalize to unknown data or has been overtrained on the training data (overfitting), comparable to “memorizing” all training questions including correct answers. This step is distinct from the definition of validation in the context of system and product development, since validation in the ML context is merely an intermediate step in the training process and not an immediate check of the final model or the system or product requirements (→ Validation in the context of system and product development). Mentioned in ISTQB – CTAI Syllabus v1.0 [137].

Term	Meaning and use
validation (in the context of system or product development)	Validation in the context of system and product development is the confirmation by the provision of objective evidence that the requirements for a specific intended use or application have been met. It is thus distinguished from verification as well as from validation in the context of ML, which aims at an optimization of hyperparameters or selection of a suitable model in the context of the training process (→ see Validation in the context of ML). Mentioned in ISO/IEC 22989:2022 [16].
validation data	In the context of ML, validation data is used to check trained ML models (→ see Validation in the context of ML) Validation data generally must not be part of the training data. Mentioned in ISO/IEC 22989:2022 [16], ISO/IEC TR 29119-11 [132].
verification	Confirmation by objective evidence that the specified requirements have been met. Note: Verification only states that a product conforms to its specification. Mentioned in ISO/IEC 22989:2022 [16].
weight	“Weights” (noun) are broadly defined as parameters of a model, usually factors that individually scale (“weight”, verb) specific entries of multidimensional inputs. In artificial neural networks, for example, weights scale the input values of an artificial neuron. In machine learning, the weights of a model are typically trained. Mentioned in ISTQB – CTAI Syllabus v1.0 [137].
work organization	Interacting work systems whose interaction achieves a specific overall result. Mentioned in DIN EN ISO 6385:2016 [235].

10
Bibliography

[1] Blind Prof. Dr. Knut; Jungmittag Prof. Dr. Andre; Mangelsdorf Dr. Axel, The Economic Benefits of Standardization. An update of the study carried out by DIN in 2000 is available at: https://www.din.de/resource/blob/79542/946e-70a818ebdaacce9705652a052b25/gesamtwirtschaftlicher-nutzen-der-normung-data.pdf (last accessed: 2022-09-26)

[2] Federal Ministry of Economic Affairs and Climate Action (BMWK), Strategie Künstliche Intelligenz der Bundesregierung (The Federal Government's Artificial Intelligence Strategy), 2018, available at www.ki-strategie-deutschland.de (last accessed: 2022-08)

[3] Bundesregierung (German Federal Government), Die entscheidende Zukunftstechnologie des 21. Jahrhunderts, 2020, available at: https://www.bundesregierung.de/breg-de/suche/fortschreibung-ki-strategie-1824340 (last accessed: 2022-09-26)

[4] European Commission, Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts Com/2021/206 Final, 2021, avail. at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206 (last accessed: 2022-08-29)

[5] Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions – Coordinated Plan on Artificial Intelligence, 2018 available at https://eur-lex.europa.eu/resource.html?uri=cellar:22ee84bb-fa04-11e8-a96d-01aa75ee d71a1.0003.02/DOC_1&format=PDF (last accessed: 2022-06-30)

[6] Independent High-Level Expert Group on AI set up by the European Commission, Policy and Investment Recommendations for Trustworthy Artificial Intelligence, 2019, avail. at: https://futurium.ec.europa.eu/en/europe-an-ai-alliance/open-library/policy-and-investment-recommendations-trustworthy-artificial-intelligence (last accessed: 2022-09-26)

[7] European Commission, WHITE PAPER, On Artificial Intelligence: A European approach to excellence and trust, 2020, avail. at: commission-white-paper-artificial-intelligence-feb2020_en.pdf (europa.eu) (last accessed: 2022-08-29)

[8] Independent High-Level Expert Group on AI set up by the European Commission, Ethics Guidelines for Trustworthy AI, 2019

[9] European Commission, Regulatory framework proposal on artificial intelligence, Juni 2022, avail. at: https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai (last accessed: 2022-08-29)

[10] IEEE 7001:2021, Standard for Transparency of Autonomous Systems

[11] IEEE 7002:2022, Standard for Data Privacy Process

[12] IEEE 7007:2021, Ontological Standard for Ethically driven Robotics and Automation Systems

[13] IEEE 7005:2021, Transparent Employer Data Governance

[14] ISO/IEC JTC1/SC 42, Artificial intelligence, avail. at: https://www.iso.org/committee/6794475.html (last accessed: 2022-09-26)

[15] ISO/IEC TR 24368:2022, Information technology – Artificial intelligence – Overview of ethical and societal concerns, avail. at: https://www.iso.org/standard/78507.html (last accessed: 2022-09-26)

[16] ISO/IEC 22989:2022, Information technology – Artificial intelligence – Artificial intelligence concepts and terminology

[17] DIN EN ISO/IEC 17065:2013, Conformity assessment – Requirements for bodies certifying products, processes and services (ISO/IEC 17065:2012); German and English version EN ISO/IEC 17065:2012

[18] DIN EN ISO/IEC 17067:2013, Conformity assessment – Fundamentals of product certification and guidelines for product certification schemes (ISO/IEC 17067:2013); German and English version EN ISO/IEC 17067:2013

[19] ISO/IEC TR 17026:2015, Conformity assessment – Example of a certification scheme for tangible products

[20] ISO/IEC TR 17028:2017, Conformity assessment – Guidelines and examples of a certification scheme for services

[21] ISO/IEC TR 17032:2019, Conformity assessment – Guidelines and examples of a scheme for the certification of processes, avail. at: https://www.iso.org/standard/29355.html (last accessed: 2022-09-26)

[22] DIN EN ISO/IEC 17021-1:2015, Conformity assessment – Requirements for bodies providing audit and certification of management systems – Part 1: Requirements

[23] Tambiama Madiega, Anne Louise Van De Pol, European Parliamentary Research Service, Artificial intelligence act and regulatory sandboxes, 2022, avail. at: https://www.europarl.europa.eu/RegData/etudes/BRIE/2022/733544/EPRS_BRI(2022)733544_EN.pdf (last accessed: 2022-09-09)

[24] ISO/IEC 23053:2022, Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML) 2022

[25] ISO/IEC DIS 23894:2022, Information technology_- Artificial intelligence_- Guidance on risk management

[26] ISO/IEC 38507:2022, Information technology – Governance of IT – Governance implications of the use of artificial intelligence by organizations, avail. at: https://www.iso.org/standard/56641.html (last accessed: 2022-09-26)

[27] ISO/IEC DIS 42001, Information Technology – Artificial intelligence – Management system

[28] ISO/IEC TR 24028:2020, Information technology – Artificial intelligence – Overview of trustworthiness in artificial intelligence

[29] ISO/IEC PRF TS 4213, Information technology – Artificial Intelligence – Assessment of machine learning classification performance

[30] ISO/IEC DIS 5338, Information technology – Artificial intelligence – AI system life cycle processes

[31] ISO/IEC CD 5339, Information Technology – Artificial Intelligence – Guidelines for AI Applications

[32] ISO/IEC CD 5392, Information technology – Artificial intelligence – Reference Architecture of Knowledge Engineering

[33] ISO/IEC DTR 5469, Artificial intelligence – Functional safety and AI systems

[34] ISO/IEC AWI TS 5471, Artificial intelligence – Quality evaluation guidelines for AI systems

[35] ISO/IEC DIS 25059:2022-07 – Software engineering_- Systems and software Quality Requirements and Evaluation (SQuaRE)_- Quality model for AI systems

[36] ISO/IEC AWI TS 6254, Information technology – Artificial intelligence – Objectives and approaches for explainability of ML models and AI systems

[37] ISO/IEC AWI TS 8200, Information technology – Artificial intelligence – Controllability of automated artificial intelligence systems

[38] ISO/IEC AWI TS 12791, Information technology – Artificial intelligence – Treatment of unwanted bias in classification and regression machine learning tasks

[39] ISO/IEC 5259 (all parts), Artificial intelligence – Data quality for analytics and machine learning (ML)

[40] ISO/IEC CD 5259-1, Artificial intelligence – Data quality for analytics and machine learning (ML) – Part 1: Overview, terminology, and examples

[41] ISO/IEC AWI 5259-2, Artificial intelligence – Data quality for analytics and machine learning (ML) – Part 2: Data quality measures

[42] ISO/IEC CD 5259-3:2022, Artificial intelligence – Data quality for analytics and machine learning (ML) – Part 3: Data quality management requirements and guidelines

[43] ISO/IEC CD 5259-4, Artificial intelligence – Data quality for analytics and machine learning (ML) – Part 4: Data quality process framework

[44] ISO/IEC AWI 5259-5, Artificial intelligence – Data quality for analytics and machine learning (ML) – Part 5: Data quality governance

[45] ISO/IEC CD 8183, Information technology – Artificial intelligence- Data life cycle framework

[46] INDEPENDENT HIGH-LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE SET UP BY THE EUROPEAN COMMISSION, A definition of AI: Main capabilities and scientific disciplines, 2019, avail. at: https://digital-strategy.ec.europa.eu/en/library/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines (last accessed: 2022-09-27)

[47] T. Schmid, W. Hildesheim, T. Holoyad, K. Schumacher, The AI Methods, Capabilities and Criticality Grid – A Three-Dimensional Classification Scheme for Artificial Intelligence Applications, 2021, avail. at: https://doi.org/10.1007/s13218-021-00736-4 (last accessed: 2022-09-26)

[48] T. Schmid; W. Hildesheim; T. Holoyad; K. Schumacher, Managing and Understanding Artificial Intelligence Solutions – The AI-Methods, Capabilities and Criticality Grid and its Value for Decision Makers, Developers and Regulators, 1st. ed, Beuth-Verlag, Berlin, 2020, https://www.beuth.de/de/publikation/kuenstliche-intelligenz-managen-und-verstehen/359390396 (last accessed: 2022-09-26)

[49] Goertzel, Ben, Perception Processing for General Intelligence: Bridging the Symbolic/Subsymbolic Gap. Artificial General Intelligence, 2012

[50] Hammer Barbara, Hitzler Pascal, Perspectives of Neural-Symbolic Integration. Studies in Computational Intelligence, 2007

[51] Russell Stuart J., Norvig Peter, Artificial Intelligence: a modern approach. Third Ed., 2014

[52] Horvitz Eric J.; Breese, John S.; Henrion, Max, Decision theory in expert systems and artificial intelligence. International Journal of Approximate Reasoning 2 (3), 1988

[53] Martin, Andreas; Hinkelmann, Knut; Gerber, Aurona; Lenat, Doug; van Harmelen, Frank; Clark, Peter, Proceedings of the AAAI 2019 Spring Symposium on Combining Machine Learning with Knowledge Engineering, 2019

[54] McGarry, Kenneth; Wermter, Stefan; MacIntyre, John, Hybrid neural systems: from simple coupling to fully integrated neural networks. Neural Computing Surveys 2 (1), 1999

[55] Méhaut, Philippe; Winch, Christopher, The European qualification framework: Skills, competences or knowledge?, 2012

[56] Bloom, Benjamin, Taxonomy of educational objectives, Vol. 1: cognitive domain, 2016

[57] Ritchie, J. Brendan, Carruthers, P, The bodily senses. In: Matthen M. (Ed.), The Oxford handbook of the philosophy of perception, 2015

[58] Macpherson, Fiona, Individuating the senses. In: Macpherson F. (ed.), The senses: classic and contemporary philosophical readings, 2011

[59] Krathwohl, David R., A revision of Bloom's taxonomy. Theory Pract 41, 2002

[60] Davidson, Donald, Essay III. In: Davidson D (ed) Essays on actions and events, 1980

[61] Shannon, Claude E., A mathematical theory of communication, 1948

[62] Luhmann, Niklas, What is communication? Communication Theory 2 (3), 1992

[63] German Institute for Standardization, the German Commission for Electrical, Electronic & Information Technologies in DIN and VDE (DKE), Publisher: Wahlster & Winterhalter, German Standardization Roadmap Artificial Intelligence, 2020

[64] IEEE 7000:2021, IEEE Standard Model Process for Addressing Ethical Concerns during System Design, avail. at: https://standards.ieee.org/ieee/7000/6781/# (last accessed: 2022-09-26)

[65] Immanuel Kant, Grundlegung zur Metaphysik der Sitten (Groundwork of the Metaphysics of Morals), 1785

[66] Enquete-Kommission (Deutscher Bundestag – Enquete-Kommission „Künstliche Intelligenz"), Bericht der Enquete-Kommission Künstliche Intelligenz – Gesellschaftliche Verantwortung und wirtschaftliche, soziale und ökologische Potenziale, 2020, avail. at: Drucksache 19/23700 (bundestag.de) (last accessed: 2022-09-26)

[67] Jobin, Anna; Ienca, Marcello; Vayena, Effi, The global landscape of AI ethics guidelines, 2019

[68] Heesen, J. et al., Ethik-Briefing. Leitfaden für eine verantwortungsvolle Entwicklung und Anwendung von KI-Systemen, 2020, avail. at: https://www.plattform-lernende-systeme.de/files/Downloads/Publikationen/AG3_Whitepaper_EB_200831.pdf (last accessed: 2022-09)

[69] Sartori, Laura; Thedorou, Andreas, A sociotechnical perspective for the future of AI: narratives, inequalities, and human control. In Ethics and Information Technology. In: Ethics and Information Technology. 24:4, 2022, avail. at: https://doi.org/10.1007/s10676-022-09624-3 (last accessed: 202209)

[70] Birhane, Abeba, Algorithmic injustice: a relational ethics approach. Patterns. 2, 2021, avail. at: https://doi.org/10.1016/j.patter.2021.100205 (last accessed: 2022-09)

[71] Bratteteig, Tone; Verne, Guri, Does AI make PD obsolete? Exploring Challenges from Artificial Intelligence to Participatory Design. In Proceedings of PDC 2018, Belgium, August 2018, 5 pages, 2018, avail. at: Does AI make PD obsolete? | Proceedings of the 15th Participatory Design Conference: Short Papers, Situated Actions, Workshops and Tutorial – Volume 2 (acm.org) (last accessed: 202209)

[72] Friedman, Batya et al., Value Sensitive Design and Information Systems. In: Doorn, Neelke; Schuurbiers, Daan; van de Poel, Ibo & Gorman, Michael E. (ed.): Early Engagement and New Technologies: Opening Up the Laboratory. Springer VS, Wiesbaden, pp. 55–96, 2013

[73] DIE ETHISCHEN LEITLINIEN DER GESELLSCHAFT FÜR INFORMATIK E. V., 2018, avail. at: Unsere Ethischen Leitlinien – Gesellschaft für Informatik e. V. (gi.de) (last accessed: 2022-09-26)

[74] Künstliche Intelligenz im Dienste der Diversität, 2022, avail. at: https://kidd-prozess.de/ (last accessed: 2022-08-12)

[75] DIN EN ISO/IEC 18045:2021, Information technology – Security techniques – Methodology for IT security evaluation (ISO/IEC 18045:2008); only on CD-ROM

[76] Arrangement on the Recognition of Common Criteria Certificates in the field of Information Technology Security, 1998, avail. at: Arrangement on the Recognition of Common Criteria Certificates (commoncriteriaportal.org) (last accessed: 2022-09-26)

[77] Gemeinsame Kriterien für die Prüfung und Bewertung der Sicherheit von Informationstechnik, avail. at: https://www.bsi.bund.de/DE/Themen/Unternehmen-und-Organisationen/Standards-und-Zertifizierung/Zertifizierung-und-Anerkennung/Zertifizierung-von-Produkten/Zertifizierung-nach-CC/IT-Sicherheiskriterien/CommonCriteria/commoncriteria_node.html (last accessed: 2022-09-26)

[78] ISO/IEC 38500:2015, Information security, cybersecurity and privacy protection – Evaluation criteria for IT security – Methodology for IT security evaluation

[79] M. Cerezo, Andrew Arrasmith, Ryan Babbush, Simon C. Benjamin, Suguru Endo, Keisuke Fujii, Jarrod R. McClean, Kosuke Mitarai, Xiao Yuan, Lukasz Cincio & Patrick J. Coles, Variational quantum algorithms, 2021, avail. at: https://www.nature.com/articles/s42254-021-00348-9 (last accessed: 2022-08-12)

[80] Prasanna Date, Davis Arthur & Lauren Pusey-Nazzaro, QUBO formulations for training machine learning models, 2021, avail. at: https://www.nature.com/articles/s41598-021-89461-4 (last accessed: 2022-0812)

[81] Bundesamt für Sicherheit in der Informationstechnik (BSI) (Federal Office for Information Security), Quantum Machine Learning in the Context of IT Security, 2022, avail. at: https://www.bsi.bund.de/DE/Service-Navi/Publikationen/Studien/QML/QML_node.html (last accessed: 20220524)

[82] Bundesamt für Sicherheit in der Informationstechnik (BSI) (Federal Office for Information Security), AI Cloud Service Compliance Criteria Catalogue (AIC4), 2021, avail. at: https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/CloudComputing/AIC4/AI-Cloud-Service-Compliance-Criteria-Catalogue_AIC4.html (last accessed: 20220510)

[83] Bundesamt für Sicherheit in der Informationstechnik, Sicherer, robuster und nachvollziehbarer Einsatz von KI Probleme, Maßnahmen und Handlungsbedarfe, 2021, Bonn, avail. at: https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/KI/Herausforderungen_und_Massnahmen_KI.pdf?__blob=publicationFile&v=6 (last accessed: 2022-08-01)

[84] D. C. Ciresan, U. Meier, J. Schmidhuber, Multi-Column Deep Neural Networks for Image Classification. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012

[85] Oliver Zendel; Markus Murschitz; Martin Humenberger; Wolfgang Herzner, „Cv-hazop: Introducing test data validation for computer vision." Proceedings of the IEEE International Conference on Computer Vision, 2015

[86] Andreas Geiger, Philip Lenz, Raquel Urtasun, Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012

[87] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, B. Schiele, The Cityscapes Dataset for Semantic Urban Scene Understanding. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016

[88] Tagiew, R.; Buder, T.; Tilly, R.; Hofmann, K.; Klotz, C., Datensätze für das autonome Fahren als Grundlage für GoA3+. In: ETR – Eisenbahntechnische Rundschau, 2021, V. 9, pp. 10–14

[89] DIN EN 50657:2017, Railways Applications – Rolling stock applications – Software on Board Rolling Stock

[90] ISO 21448:2022, Road vehicles – Safety of the intended functionality

[91] ISO/IEC TR 24029-1:2021, Artificial Intelligence (AI) – Assessment of the robustness of neural networks – Part 1: Overview

[92] ISO/IEC DIS 24029-2, Artificial intelligence (AI) – Assessment of the robustness of neural networks – Part 2: Methodology for the use of formal methods

[93] C. Hasterok et al., PAISE – The process model for AI systems engineering, 2021. avail. at https://www.ki-engineering.eu/content/dam/iosb/ki-engineering/downloads/PAISE(R)_Whitepaper_english.pdf (last accessed 2023-01-10)

[94] Trapp, Mario; Schneider, Daniel; Weiss, Gereon, Towards safety-awareness and dynamic safety management. 2018 14th European Dependable Computing Conference (EDCC). IEEE

[95] Ruf, M. et al., Comparison of local vs. global optimization for trajectory planning in automated driving. 10. Workshop Fahrerassistenzsysteme, 2015

[96] Prof. Dr. Dr. Udo Di Fabio, Prof. Dr. Dr. h.c. Manfred Broy, Renata Jungo Brüngger, Dr. Ulrich Eichhorn, Prof. Dr. Armin Grunwald, Prof. Dr. Dirk Heckmann, Prof. Dr. Dr. Eric Hilgendorf, Prof. Dr. rer. Nat. Dr.-Ing. E. h. Henning Kagermann, Weihbischof Dr. Dr. Anton Losinger, Prof. Dr. Dr. Matthias Lutz-Bachmann, Prof. Dr. Christoph Lütge, Dr. August Markl, Klaus Müller, Kay Nehm, Bericht der Ethikkommission zum automatisierten und vernetzten Fahren, BMVI, 2017

[97] N. Heide, A Step towards Explainable Artificial Neural Networks in Image Processing by Dataset Assessment. Forum Bildverarbeitung, 2020

[98] DIN SPEC 13266:2020, Guideline for the development of deep learning image recognition systems

[99] ISO/IEC Guide 51:2014, Safety aspects – Guidelines for their inclusion in standards, avail. at: https://www.beuth.de/de/technische-regel/iso-iec-guide-51/205060593 (last accessed: 2022-09-26)

[100] DIN CLC IEC/TR 63069:2021, Industrial-process measurement, control and automation – Framework for functional safety and security (IEC/TR 63069:2019), avail. at: https://www.beuth.de/de/norm/din-clc-iec-tr-63069/332535627 (last accessed: 2022-09-26)

[101] DIN EN 61508-1:2011, VDE 0803-1:2011, Functional safety of electrical/electronic/programmable electronic safety-related systems – Part 1: General requirements (IEC 61508-1:2010)

[102] DIN EN 61508-2:2011, VDE 0803-2:2011-02, Functional safety of electrical/electronic/programmable electronic safety-related systems – Part 2: Requirements for electrical/electronic/programmable electronic safety-related systems (IEC 61508-2:2010)

[103] DIN EN 615083:2011, VDE 0803-3:2011, Functional safety of electrical/electronic/programmable electronic safety-related systems – Part 3: Software requirements (IEC 61508-3:2010)

[104] NIST Special Publication 1011-I-2.0:2008, Autonomy Levels for Unmanned Systems (ALFUS) Framework, Volume I – Terminology

[105] VDE-AR-E 2842-61-2 Anwendungsregel:2021-06, Entwicklung und Vertrauenswürdigkeit von autonom/kognitiven Systemen

[106] ISO 21815-1:2022, Earth-moving machinery – Collision warning and avoidance – Part 1: General requirements

[107] ISO/TS 21815-2:2021, Earth-moving machinery – Collision warning and avoidance – Part 2: Onboard J1939 communication interface

[108] ISO/DIS 21815-3:2022- Draft, Earth-moving machinery – Collision warning and avoidance – Part 3: Risk area and risk level – Forward/reverse motion

[109] DIN EN ISO 13849-1:2016, Safety of machinery – Safety-related parts of control systems – Part 1: General principles for design (ISO 138491:2015)

[110] ISO/AWI PAS 8800, Road Vehicles – Safety and artificial intelligence

[111] DIN EN ISO 138491:2021 – Draft, Safety of machinery – Safety-related parts of control systems – Part 1: General principles for design (ISO/DIS 13849-1.2:2021)

[112] DIN EN ISO 25119-1:2022-02 – Draft, Tractors and machinery for agriculture and forestry – Safety-related parts of control systems – Part 1: General principles for design and development (ISO 25119-1:2018)

[113] ISO/IEC/IEEE DIS 150263:2022, Systems and software engineering – Systems and software assurance – Part 3: System integrity levels

[114] ISO/IEC/IEE 150261:2019, International Standard – Systems and software engineering–Systems and software assurance – Part 1: Concepts and vocabulary, avail. at: https://ieeexplore.ieee.org/document/8657410 (last accessed: 2022-09-26)

[115] EU Observatory for ICT Standardisation, avail. at: https://www.standict.eu/euos (last accessed: 2022-09-26)

[116] EU Observatory for ICT Standardisation, Report of TWG AI: Landscape of AI Standards, avail. at: https://zenodo.org/record/5011179#.YhvgLOjMK5c (last accessed: 2022-09-26)

[117] DIN SPEC 92001-3, Artificial Intelligence – Life Cycle Processes and Quality Requirements – Part 3: Explainability

[118] European Union Agency for Cybersecurity, About ENISA – The European Union Agency for Cybersecurity, 2022, avail. at: https://www.enisa.europa.eu/about-enisa/about-enisa-the-european-union-agency-for-cybersecurity (last accessed: 2022-09-22)

[119] European Union Agency for Cybersecurity, Securing Machine Learning Algorithms, 2021, avail. at: https://www.enisa.europa.eu/publications/securing-machine-learning-algorithms (last accessed: 2022-09-22)

[120] Poretschkin, Maximilian; Schmitz, Anna; Akila, Maram; Adilova, Linara; Becker, Daniel; Cremers, Armin B.; Hecker, Dirk; Houben, Sebastian; Mock, Michael; Rosenzweig, Julia; Sicking, Joachim; Schulz, Elena; Voss, Angelika; Wrobel, Stefan, Leitfaden zur Gestaltung vertrauenswürdiger künstlicher Intelligenz (KI-Prüfkatalog), 2021, Sankt Augustin

[121] ISO/IEC AWI 27090, Cybersecurity – Artificial Intelligence – Guidance for addressing security threats and failures in artificial intelligence systems

[122] ISO/IEC 27034-1:2011, Information technology – Security techniques – Application security – Part 1: Overview and concepts, avail. at: https://www.iso.org/standard/44378.html (last accessed: 2022-09-26)

[123] ISO/IEC 27034-2:2015, Information technology – Security techniques – Application security – Part 2: Organization normative framework

[124] ISO/IEC 27034-3:2018, Information technology – Security techniques – Application security – Part 3: Application security management process

[125] ISO/IEC 27034-5:2017, Information technology – Security techniques – Application security – Part 5: Protocols and application security controls data structure

[126] ISO/IEC 27034-6:2016, Information technology – Security techniques – Application security – Part 6: Case studies

[127] ISO/IEC 27034-7:2018, Information technology – Security techniques – Application security – Part 7: Assurance prediction framework

[128] DIN EN ISO/IEC 27701:2021, Security techniques – Extension to ISO/IEC 27001 and ISO/IEC 27002 for privacy information management – Requirements and guidelines(ISO/IEC 27701:2019)

[129] Second Data Protection Adaptation and Implementation Act EU 2016/679 and implementing Directive (EU) 2016/680 (Zweites Datenschutz-Anpassungs- und Umsetzungsgesetz EU – 2. DSAnpUG-EU), 2019, https://www.bgbl.de/xaver/bgbl/media/FEE38353527D8A83E26346A53BE44BD7/bgbl119s1626_77927.pdf (last accessed: 2022-09-26)

[130] DIN EN ISO/IEC 27037:2016, Information technology – Security techniques – Guidelines for identification, collection, acquisition and preservation of digital evidence (ISO/IEC 27037:2012)

[131] DIN EN ISO/IEC 27000 series, Continuously supplemented ISO standard series on the subject of information security, excerpt: DIN EN ISO/IEC 27000:2014 Information technology – Security techniques – Information security management systems – Overview and vocabulary, DIN EN ISO/IEC 27001:2022 Information security, cybersecurity and privacy protection – Information security management systems – Requirements, DIN EN ISO/IEC 27002:2022 Information security, cybersecurity and privacy protection – Information security controls

[132] ISO/IEC AWI TS 29119-11, Information technology – Artificial intelligence – Testing for AI systems – Part 11

[133] DIN EN ISO/IEC 29100:2020, Information technology – Security techniques – Privacy framework (ISO/IEC 29100:2011, including Amd 1:2018), avail. at: https://www.beuth.de/de/norm/din-en-iso-iec-29100/325198919 (last accessed 2022-09-26)

[134] DIN EN ISO/IEC 29134:2020, Information technology – Security techniques – Guidelines for privacy impact assessment (ISO/IEC 29134:2017)

[135] DIN EN ISO/IEC 29151:2022, Information technology – Security techniques – Code of practice for personally identifiable information protection (ISO/IEC 29151:2017)

[136] Bitkom Bundesverband Informationswirtschaft, Telekommunikation und neue Medien e. V. Machine Learning und die Transparenzanforderungen der DS-GVO Leitfaden, 2018, https://www.bitkom.org/sites/default/files/file/import/180926-Machine-Learning-und-DSGVO.pdf (last accessed: 2022-09-26)

[137] ISTQB, Lehrplan zum Certified Tester AI Testing, avail. athttps://www.istqb.org/certifications/artificial-inteligence-tester (last accessed: 2022-09-26)

[138] ISO/IEC CD TR 27563:2022, Security and privacy in artificial intelligence use cases, avail. at: https://www.iso.org/standard/80396.html (last accessed: 2022-09-26)

[139] U.S. Food & Drug Administration, Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) – Discussion Paper and Request for Feedback, 2019, avail. at: https://www.fda.gov/media/122535/download (last accessed: 2022-08-17)

[140] Francesco Croce, Matthias Hein, Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks, 2020, avail. at: https://proceedings.mlr.press/v119/croce20b.html (last accessed: 2022-09-26)

[141] Mock, M.; Scholz, S.; Blank, F.; Hüger; F., Rohatschek, A.; Schwarz, L.; Stauner, T., SAFECOMP Workshops, Springer, An Integrated Approach to a Safety Argumentation for AI-Based Perception Functions in Automated Driving, 2021

[142] M. Mock, A. Schmitz, Fraunhofer IAIS, Management System Support for Trustworthy Artificial Intelligence, 2021

[143] Lernende Systeme – Die Plattform für Künstliche Intelligenz, Lernende Systeme im Gesundheitswesen. Grundlagen, Anwendungsszenarien und Gestaltungsoptionen, 2019, avail. at: https://www.plattform-lernende-systeme.de/files/Downloads/Publikationen/AG6_Lernende_Systeme_im_Gesundheitswesen_web_final.pdf (last accessed: 2022-09-26)

[144] DECISION No 768/2008/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 9 July 2008 on a common framework for the marketing of products, and repealing Council Decision 93/465/EEC, avail. at https://eur-lex.europa.eu/legal-content/de/ALL/?uri=CELEX:32008D0768 (last accessed: 2022-09-26)

[145] REGULATION (EC) No 765/2008 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 9 July 2008 setting out the requirements for accreditation and market surveillance relating to the marketing of products and repealing Regulation (EEC) No 339/93

[146] Directive 2006/42/EC of the European Parliament and of the Council of 17 May 2006 on machinery and amending Directive 95/16/EC

[147] DIN EN ISO/IEC 17000:2020, Conformity assessment – Vocabulary and general principles (ISO/IEC 17000:2020)

[148] ISO/IEC/IEEE 12207:2017, Systems and software engineering – Software life cycle processes, avail. at: https://www.iso.org/standard/63712.html (last accessed: 2022-09-26)

[149] ISO/IEC CD 5394:2021, Information Technology – Artificial intelligence – Management System

[150] ISO/IEC WD TS 24462:2022, Ontology for ICT Trustworthiness Assessment

[151] ISO/DIS 24089, Road vehicles – Software update engineering

[152] ISO/IEC 25010:2011, Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – System and software quality models, avail. at:https://www.iso.org/standard/35733.html (last accessed: 2022-09-26)

[153] N. Beck, C. Martens, K.H. Sylla, D. Wegener, A. Zimmermann, ZUKUNFTSSICHERE LÖSUNGEN FÜR MASCHINELLES LERNEN, MACHINE LEARNING OPERATIONS (MLOPS) – PROZESSE FÜR ENTWICKLUNG, INTEGRATION UND BETRIEB, 2021

[154] S. Beck, Plattform lernende Systeme, Künstliche Intelligenz und Diskriminierung. Herausforderungen und Lösungsansätze. Plattform Lernende Systeme (Hrsg.), 2019, avail. at: https://www.plattform-lernende-systeme.de/files/Downloads/Publikationen/AG3_Whitepaper_250619.pdf (last accessed: 2022-08-24)

[155] DIN EN ISO/IEC 17024:2012, Conformity assessment – General requirements for bodies operating certification of persons (ISO/IEC 17024:2012)

[156] DIN EN ISO/IEC 17025:2018, General requirements for the competence of testing and calibration laboratories (ISO/IEC 17025:2017)

[157] DIN EN ISO/IEC 17020:2012, Conformity assessment – Requirements for the operation of various types of bodies performing inspection (ISO/IEC 17020:2012)

[158] DIN EN ISO/IEC 17029:2020, Conformity Assessment – General principles and requirements for validation and verification bodies (ISO/IEC 17029:2019)

[159] DIN EN ISO/IEC 17011:2018, Conformity assessment – Requirements for accreditation bodies accrediting conformity assessment bodies (ISO/IEC 17011:2017)

[160] DIN ISO 31000:2018, Risk management – Guidelines (ISO 31000:2018)

[161] ISO/IEC 27005:2018, Information technology – Security techniques – Information security risk management

[162] DIN SPEC 92001-1:2019, Artificial Intelligence – Life Cycle Processes and Quality Requirements – Part 1: Quality Meta Model, avail. at: https://www.beuth.de/de/technische-regel/din-spec-92001-1/303650673 (last accessed: 2022-09-26)

[163] REGULATION (EU) 2019/881 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 17 April 2019 on ENISA (the European Union Agency for Cybersecurity) and on information and communications technology cybersecurity certification and repealing Regulation (EU) No 526/2013 (Cybersecurity Act) (Text with EEA relevance)

[164] REGULATION (EU) No 526/2013 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 21 May 2013 concerning the European Union Agency for Network and Information Security (ENISA) and repealing Regulation (EC) No 460/2004 (Text with EEA relevance)

[165] KI-LOK sicher KI für die Schiene, avail. at: https://ki-lok.itpower.de/ (last accessed: 2022-09-26)

[166] GDPR, General Data Protection Regulation 2016, avail. at: https://eur-lex.europa.eu/legal-content/DE-EN/TXT/?-from=DE&uri=CELEX%3A32016R0679 (last accessed: 2022-09-26)

[167] ProdSG, Gesetz über die Bereitstellung von Produkten auf dem Markt (Produktsicherheitsgesetz) (German Product Safety Act), avail. at: https://www.gesetze-im-internet.de/prodsg_2021/ (last accessed: 2022-09-26)

[168] The Fraunhofer Institute for Open Communication Systems FOKUS, Industrial Grade Machine Learning for Enterprises, 2022, avail. at: https://iml4e.org/ (last accessed: 2022-09-26)

[169] REGULATION (EU) No 1025/2012 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 25 October 2012 on European standardisation, amending Council Directives 89/686/EEC and 93/15/EEC and Directives 94/9/EC, 94/25/EC, 95/16/EC, 97/23/EC, 98/34/EC, 2004/22/EC, 2007/23/EC, 2009/23/EC and 2009/105/EC of the European Parliament and of the Council and repealing Council Decision 87/95/EEC and Decision No 1673/2006/EC of the European Parliament and of the Council (Text with EEA relevance)

[170] Horizontal Harmonization Committee, EA-1/06 A-AB:2022, EA Multilateral Agreement. Criteria for signing. Policy and procedures for development, 2022, avail. at: https://european-accreditation.org (last accessed: 2022-09-26)

[171] International Accreditation Forum, IAF PR4: 2015, Structure of the IAF MLA and List of IAF Endorsed Normative Documents, avail. at: https://european-accreditation.org (last accessed: 2022-09-26)

[172] Artificial Intelligence and Data Act, avail. at: https://www.osler.com/en/resources/regulations/2022/government-of-canada-s-artificial-intelligence-and-data-act-brief-overview (last accessed: 2022-09-26)

[173] T. Hagendorff, The Ethics of AI Ethics. An Evaluation of Guidelines, Minds and Machines, p. 122, 2020, avail. at: https://arxiv.org/pdf/1903.03425.pdf (last accessed: 2022-09-26)

[174] Zweig, K. A., Krafft, T. D., Klingel, A., & Park, E., Sozioinformatik: ein neuer Blick auf Informatik und Gesellschaft. Carl Hanser Verlag GmbH Co KG, 2021

[175] Schlick, Christopher; Bruder, Ralph; Luczak, Holger, Arbeitswissenschaft. Springer-Verlag Berlin Heidelberg. 3. Auflage, 2010

[176] Suchman, L., Human–Machine Reconfigurations. Plans and Situated Actions, 2nd Edition. Cambridge: Cambridge University Press, 2007

[177] Emery, Frederick E./Trist, Eric L., Socio-technical Systems. In: Frederick E. Emery (ed.): Systems Thinking. Harmondsworth, 1969, pp. 281–295

[178] Sydow, J., Der soziotechnische Ansatz der Arbeits- und Organisationsgestaltung, 1985

[179] Lee, J.D., Wickens, C.D., Liu, Y. & Ng Boyle, L., Designing for People: An Introduction to Human Factors Engineering. Charlston: CreateSpace, 2017

[180] DIN EN 614-1:2009, Safety of machinery – Ergonomic design principles – Part 1: Terminology and general principles

[181] DIN EN 614-2:2008, Safety of machinery – Ergonomic design principles – Part 2: Interactions between the design of machinery and work tasks

[182] DIN CEN/TR 614-3:2011, Safety of machinery – Part 3: Ergonomic principles for the design of mobile machinery

[183] DIN EN ISO 9241-210:2020, Ergonomics of human-system interaction – Part 210: Human-centred design for interactive systems, avail. at: https://www.iso.org/standard/77520.html (last accessed: 2022-09-26)

[184] Raisch S. & Krakowski S., Artificial Intelligence and Management: The Automation-Augmentation Paradox. Academy of Management Review, 2020

[185] Floridi L. & Sanders J., On the Morality of Artificial Agents. Minds and Machines, 14, 349–379, 2004

[186] MAKARIUS, E. E., MUKHERJEE, D., FOX, J. D. & FOX, A. K., Rising with the machines: A sociotechnical framework for bringing artificial intelligence into the organization. Journal of Business Research, 120, 262–273, 2020

[187] ŁAPIŃSKA, J., ESCHER, I., GÓRKA, J., SUDOLSKA, A. & BRZUSTEWICZ, Employees' Trust in Artificial Intelligence in Companies: The Case of Energy and Chemical Industries in Poland. Energies, 14, 1942, 2021

[188] S. Kugele, A. Petrovska, and I. Gerostathopoulos, “Towards a Taxonomy of Autonomous Systems,” in Software Architecture, vol. 12857, S. Biffl, E. Navarro, W. Löwe, M. Sirjani, R. Mirandola, and D. Weyns, Eds. Cham: Springer International Publishing, 2021, pp. 37–45. Doi: 10.1007/978-3-030-86044-8_3, 2021

[189] Weyer, Johannes, Die Kooperation menschlicher Akteure und nicht-menschlicher Agenten. Ansatzpunkte einer Soziologie hybride Systeme (= Arbeitspapier 16 der Wirtschafts- und Sozialwissenschaftlichen Fakultät), Dortmund: Technische Universität Dortmund, 2006

[190] Elisabeth André & Wilhelm Bauer, Kompetenzentwicklung für Künstliche Intelligenz – Veränderungen, Bedarfe und Handlungsoptionen, Whitepaper aus der Plattform Lernende Systeme, München, 2021, avail. at: https://doi.org/10.48669/pls_2021-2 (last accessed: 2022-09-26)

[191] HÖDDINGHAUS, M., SONDERN, D. & HERTEL, G., The automation of leadership functions: Would people trust decision algorithms? Comput. Hum. Behav., 116, 106635, 2021

[192] HUANG, M.-H., RUST, R. & MAKSIMOVIC, V., The Feeling Economy: Managing in the Next Generation of Artificial Intelligence (AI). California Management Review, 61, 43-65, 2019

[193] BRYNJOLFSSON, E., MITCHELL, T. & ROCK, D., What Can Machines Learn, and What Does It Mean for Occupations and the Economy? AEA Papers and Proceedings, 108, 43–47, 2018

[194] Moray, in J. Noyes & M. Bransby (Eds.), People in control. Human factors in control room design (101–115), Human and machines: Allocation of functions, 1989

[195] W. Hacker; P Sachse, Allgemeine Arbeitspsychologie. Psychische Regulation von Tätigkeiten. Göttingen: Hogrefe, 2014

[196] Watzlawick, P., Beavin, J.H., Jackson, D.D., Menschliche Kommunikation: Formen, Störungen, Paradoxien. Bern: Huber, 2011

[197] Cherns A., The Principles of Sociotechnical Design. Human Relations 29 (8), 783–792. DOI: https://doi.org/10.1177/001872677602900806, 1976

[198] Cherns A., Principles of sociotechnical design revisited. Human Relations, 40 (3), 153–161. DOI: https://doi.org/10.1177/001872678704000303, 1987

[199] Ulich, Eberhard, Arbeitssysteme als Soziotechnische Systeme – eine Erinnerung. Journal Psychologie des Alltagshandelns / Psychology of Everyday Activity, Vol. 6 / No. 1, ISSN 1998-9970, p. 4–12, 2013

[200] BENJAMIN, RUHA, Race after Technology. Abolitionist Tools for the New Jim Code. Cambridge/Medford: Polity Press, 2019

[201] Pentenrieder, Annelie; Weber, Jutta, Lucy Suchman (geb. 1951). In: Heßler, Martina & Liggieri, Kevin (ed.). Technikanthropologie: Handbuch für Wissenschaft und Studium. Nomos: Baden-Baden. pp. 215–224, 2020

[202] Dr. Norbert Huchler et al., Kriterien für die Mensch-Maschine-Interaktion bei KI. Ansätze für die menschengerechte Gestaltung in der Arbeitswelt, 2020, avail. at: https://www.plattform-lernende-systeme.de/files/Downloads/Publikationen/AG2_Whitepaper2_220620.pdf (last accessed: 2022-09-26)

[203] Sascha Stowasser & Oliver Suchy et al. (ed.), Einführung von KI-Systemen in Unternehmen. Gestaltungsansätze für das Change-Management. Whitepaper aus der Plattform Lernende Systeme, München 2020

[204] Zink, K. J., Soziotechnische Ansätze. In H. Luczak & M. Volpert (ed.), Handbuch Arbeitswissenschaft (pp. 74–77). Stuttgart: Schäffer-Poeschel, 1997

[205] Ulich, Eberhard, Arbeitspsychologie, 2011

[206] Bergmann, B. & Richter, P. (ed.), Die Handlungsregulationstheorie: Von der Praxis einer Theorie. Göttingen: Hogrefe, 1994

[207] Bendel A. & Latniak E., Soziotechnisch – agil – lean: Konzepte und Vorgehensweisen für Arbeits- und Organisationsgestaltung in Digitalisierungsprozessen. Gr Interakt Org (2020) 51:285–297, avail. at: https://doi.org/10.1007/s11612-020-00528-8 (last accessed: 2022-09-26)

[208] KAHNEMAN, D. & TVERSKY, Intuitive prediction: Biases and corrective procedures. In: TVERSKY, A., KAHNEMAN, D. & SLOVIC, P. (eds.) Judgment under Uncertainty: Heuristics and Biases. Cambridge: Cambridge University Press, 1982

[209] WACHTER, S. & MITTELSTADT, B., A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI Columbia Business Law Review, 2, 2019

[210] Shin, D. and Y.J. Park, Role of fairness, accountability, and transparency in algorithmic affordance [online]. Computers in Human Behavior, 98, 277-284. ISSN 07475632, avail. at: http://doi:10.1016/j.chb.2019.04.019, 2019

[211] Ostrom, A.L., D. Fotheringham and M.J. Bitner, Customer Acceptance of AI in Service Encounters: Understanding Antecedents and Consequences. In: P.P. Maglio, C.A. Kieliszewski, J.C. Spohrer, K. Lyons, L. Patrício and Y. Sawatani, ed. Handbook of Service Science, Volume II. Cham: Springer International Publishing, pp. 77–103. ISBN 978-3-319-98511-4, 2019

[212] Jeffrey Dastin, Amazon scraps secret AI recruiting tool that showed bias against women, 2018, avail. at: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G (last accessed: 2022-08-12)

[213] Simbeck K., Diskriminiert durch Künstliche Intelligenz – Ethische Aspekte beim Einsatz von analytischen, datengetriebenen Verfahren im Personalmanagement. In: Zukunft der Arbeit. Soziotechnische Gestaltung der Arbeitswelt im Zeichen von „Digitalisierung“ und „Künstlicher Intelligenz“, pp. 199–210, 2020

[214] BONNEFON, Jean-François; SHARIFF, Azim; RAHWAN, Iyad., The trolley, the bull bar, and why engineers should care about the ethics of autonomous cars [point of view]. Proceedings of the IEEE, 2019, 107. Yr., No. 3, pp. 502–504, 2019

[215] Alexander G. Mirnig and Alexander Meschtscherjakov, Trolled by the Trolley Problem: On What Matters for Ethical Decision Making in Automated Vehicles. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI ‘19). Association for Computing Machinery, New York, NY, USA, Paper 509, 1–10. https://doi.org/10.1145/3290605.3300739, 2019

[216] European Parliament and the Council, Directive 2006/42/EC on machinery, and amending Directive 95/16/EC, Mai 2006, avail. at: https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2006:157:0024:0086:en:PDF (last accessed: 2022-09-26)

[217] DIRECTIVE 2009/127/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 21 October 2009 amending Directive 2006/42/EC with regard to machinery for pesticide application (Text with EEA relevance), avail. at: https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2009:310:0029:0033:de:PDF (last accessed: 2022-09-26)

[218] Deutsche Gesetzliche Unfallversicherung, Industrie 4.0: Herausforderungen für die Prävention – Positionspapier der gesetzlichen Unfallversicherung. Positionspapier 02/2017, avail. at: https://www.dguv.de/medien/inhalt/praevention/arbeitenvierpunktnull/pospap-2-2017.pdf (last accessed: 2022-08-12)

[219] Karl Ludwig von Bertalanffy, General System Theory. In: Biologia Generalis. 1/1949, pp. 114–129

[220] Karl Ludwig von Bertalanffy, The Theory of Open Systems in Physics and Biology. In: Science. Vol. 111, 1950, pp. 23–29

[221] Talcott Parsons, The Social System. Free Press, New York, 1951

[222] Luhmann, Niklas, Soziale Systeme. 1. Auflage. Suhrkamp, Frankfurt am Main 1984, ISBN 3-518-28266-2

[223] Rohde, Friederike; Wagner, Josephin; Reinhard, Philipp; Petschow, Ulrich; Meyer, Andreas; Voß, Marcus; Mollen, Anne, Nachhaltigkeitskriterien für künstliche Intelligenz – Entwicklung eines Kriterien- und Indikatorensets für die Nachhaltigkeitsbewertung von KI-Systemen entlang des Lebenszyklus, 2021, avail. at: https://www.ioew.de/publikation/nachhaltigkeitskriterien_fuer_kuenstliche_intelligenz (last accessed: 2022-07-12)

[224] Selma Muhammad, The Fairness Handbook, 2022, avail. at: https://www.amsterdamintelligence.com/resources/the-fairness-handbook (last accessed: 2022-09-26)

[225] Saleiro, P., Rodolfa, K. T., & Ghani, R., Dealing with Bias and Fairness in Data Science Systems: A Practical Hands-on Tutorial, 2020, avail. at: https://doi.org/10.1145/3394486.3406708 (last accessed: 2022-09-26)

[226] Buolamwini, J., & Gebru, T., Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Proceedings of Machine Learning Research, 81, 77–91, 2018, avail. at: http://proceedings.mlr.press/v81/buolamwini18a.html (last accessed: 2022-09-26)

[227] Silberg, J., & Manyika, J., Notes from the AI frontier: Tackling bias in artificial intelligence (and in humans). Mckinsey Global Institute, 1–8, 2019, avail. at: https://www.mckinsey.com/featured-insights/artificial-intelligence/tackling-bias-in-artificial-intelligence-and-in-humans# (last accessed: 2022-09-26)

[228] Weerts, H. J. P., An Introduction to Algorithmic Fairness. 1–18, 2021, avail. at: http://arxiv.org/abs/2105.05595 (last accessed: 2022-09-26)

[229] Simonsen, J., & Robertson, T., Routledge international handbook of participatory design (Vol. 711). New York: Routledge, 2013

[230] Puntschuh, M., & Fetic, L., Praxisleitfaden zu den Algo.Rules. Orientierungshilfen für Entwickler:innen und ihre Führungskräfte. Bertelsmann Stiftung, Gütersloh, 2020, avail. at: https://doi.org/10.11586/2020029 (last accessed: 2022-09-26)

[231] VDE, Bertelsmann Stiftung (ed.), From Principles to Practice – An interdisciplinary framework to operationalise AI ethics. AI Ethics Impact Group, VDE Association for Electrical Electronic & Information Technologies e. V., Bertelsmann Stiftung, 1–56, 2020, avail. at: https://doi.org/10.11586/2020013 (last accessed: 2022-09-26)

[232] Puntschuh, M., & Fetic, L., Handreichung für die digitale Verwaltung. Algorithmische Assistenzsysteme gemeinwohlorientiert gestalten. Bertelsmann Stiftung, Gütersloh, 2020, avail. at: https://doi.org/10.11586/2020060 (last accessed: 2022-09-26)

[233] Mökander, J., Sheth, M., Watson, D. W., & Floridi, L., Models for Classifying AI Systems: the Switch, the Ladder, and the Matrix. In 2022 ACM Conference on Fairness, Accountability, and Transparency (FaccT'22), June 21–24, 2022, Seoul, Republic of Korea. ACM, New York, NY, USA, 1 page, avail. at: https://doi.org/10.1145/3531146.3533162 (last accessed: 2022-09-26)

[234] IG Metall Vorstand (2019), Ressort Zukunft der Arbeit (ed.), KOMPASS DIGITALISIERUNG. Eine Gestaltungshilfe für gute digitale Arbeit. 1. Auflage 2019

[235] DIN EN ISO 6385:2016, Ergonomics principles in the design of work systems (ISO 6385:2016)

[236] Ulich, Eberhard, Mensch, Technik, Organisation: ein europäisches Produktionskonzept. In O. Strohm & E. Ulich (ed), Unternehmen arbeitspsychologisch bewerten (pp. 5–17). Schriftenreihe Mensch, Technik, Organisation, Band 10 (ed. E. Ulich). Zürich: vdf Hochschulverlag, 1997

[237] Bundesanstalt für Arbeitsschutz und Arbeitsmedizin (ed.), Rechtliche Rahmenbedingungen für die Bereitstellung autonomer und KI-Systeme. Bericht, F 2432. 1. Auflage 2021, avail. at: https://www.baua.de/DE/Angebote/Publikationen/Berichte/F2432.pdf?__blob=publicationFile&v=5 (last accessed: 2022-09-26)

[238] ISO/IEC WI 12792, Information technology – Artificial intelligence – Transparency taxonomy of AI systems

[239] DIN EN ISO 26800:2011, Ergonomics – General approach, principles and concepts (ISO 26800:2011)

[240] DIN SPEC 92001-2:2020, Artificial Intelligence – Life Cycle Processes and Quality Requirements – Part 2: Robustness

[241] VDI/VDE-MT 7100 – Draft, Learning-friendly work design – Goals, benefits, terms and definitions

[242] VDE SPEC 90012:2022, VCIO based description of systems for AI trustworthiness characterization

[243] DIN EN ISO 11064:2001, Ergonomic design of control centres – Part 1: Principles for the design of control centres (ISO 11064-1:2000)

[244] DIN EN 894-1:2009, Safety of machinery – Ergonomics requirements for the design of displays and control actuators – Part 1: General principles for human interactions with displays and control actuators

[245] DIN EN ISO 9241-11:2018, Ergonomics of human-system interaction – Part 11: Usability: Definitions and concepts (ISO 9241-11:2018)

[246] DIN EN ISO 9241-110:2020, Ergonomics of human-system interaction – Part 110: Interaction principles (ISO 9241-110:2020)

[247] ISO/IEC 29138-1:2018, Information technology – User interface accessibility – Part 1: User accessibility needs

[248] VDI/VDE 3850-1:2014, Development of usable user interfaces for technical plants – Concepts, principles and fundamental recommendations

[249] DIN EN ISO 9241-112:2017, Ergonomics of human-system interaction – Part 112: Principles for the presentation of information (ISO 9241-112:2017)

[250] Hacker, Software-Ergonomie: Gestalten rechnergestützter geistiger Arbeit?! In W. Schönpflug & M. Wittstock (ed.), Software-Ergonomie, 87. Nützen Informationssysteme dem Benutzer? (pp. 31–54). Stuttgart: Teubner, 1987

[251] Hacker, Software-Gestaltung als Arbeitsgestaltung. In K.-P. Fälmrich (ed.), Software-Ergonomie, State of the Art (pp. 29-42). München: Oldenburg, 1987

[252] Böde, Eckard; Hartmann, Ernst A.; Lüdtke, Andreas (u. a.), Mensch-Technik-Interaktion. Leitfaden für Hersteller und Anwender. Vol 3. Ed.: Bundesministerium für Wirtschaft und Technologie (BMWi), 2013, avail. at: https://www.digitale-technologien.de/DT/Redaktion/DE/Downloads/Publikation/autonomik-Leitfaden3.pdf?__blob=publicationFile&v=3 (last accessed: 2022-09-26)

[253] ISO 9241-2, Ergonomic requirements for office work with visual display terminals (VDTs); part 2: Guidance on task requirements

[254] Ulich, Eberhard, Aufgabengestaltung. In H. Schmidt & U. Kleinbeck (ed.), Enzyklopädie der Psychologie (Arbeitspsychologie) (pp. 581–622), 2010

[255] Deutsche Gesetzliche Unfallversicherung, Softwareergonomie, 2021

[256] Sheridan, TB, Humans and Automation – System Design and Research Issues. Wiley, New York, 2002

[257] P.M. Fitts, Human engineering for an effective air-navigation and traffic-control system. Washington, DC: National Research Council, 1951

[258] Kraiss KF, Schmidtke H, Funktionsteilung Mensch-Machine. In: Bundesamt für Wehrtechnik und Beschaffung (ed.), Handbuch der Ergonomie. Carl Hanser, München, 2002

[259] Jordan, N., Allocation of functions between man and machines in automated systems. Journal of Applied Psychology 47, 161–165, 1963

[260] Dekker, S., Woods, D., MABA-MABA or Abracadabra? Progress on Human–Automation Coordination . Cognition, Technology & Work 4, 240–244, 2002

[261] P. A. Hancock, Automation: how much is too much? Ergonomics 57, 449–454, 2014

[262] L. Bainbridge, Ironies of Automation. In: Rasmussen J, Duncan K, Leplat J (ed.) New Technology and Human Error, John Wiley, New York, 1987

[263] DIN EN ISO 9001:2015, Quality management systems – Requirements (ISO 9001:2015)

[264] DIN EN ISO 9000:2015, Quality management systems – Fundamentals and vocabulary (ISO 9000:2015)

[265] DIN EN ISO 14001:2015, Environmental management systems – Requirements with guidance for use (ISO 14001:2015)

[266] DIN EN ISO 50001:2018, Energy management systems – Requirements with guidance for use (ISO 50001:2018)

[267] DIN ISO 45001:2018, Occupational health and safety management systems – Requirements with guidance for use (ISO 45001:2018)

[268] DIN ISO 21500:2016, Project, programme and portfolio management – Context and concepts

[269] DIN 69901 (all parts), Project management – Project management systems

[270] DIN 69909 (all parts), Multi Project Management – Management of project portfolios, programmes and projects

[271] DIN EN ISO 27500:2017, The human-centred organization – Rationale and general principles (ISO 27500:2016)

[272] „Forum Soziale Technikgestaltung“, Projekt „Der mitbestimmte Algorithmus“, Projekt PROTIS-BIT, Kriterien zur Gestaltung algorithmischer Steuerungs- und Entscheidungssysteme, Schröter: Welf: Der mitbestimmte Algorithmus. Gestaltungskompetenz für den Wandel der Arbeit, 2019, avail. at: www.blog-zukunft-der-arbeit.de (last accessed: 2022-09-26)

[273] Jessica Heesen, Jörn Müller-Quade, Stefan Wrobel et al., Kritikalität von KI-Systemen in ihren jeweiligen Anwendungskontexten – Ein notwendiger, aber nicht hinreichender Baustein für Vertrauenswürdigkeit. Whitepaper aus der Plattform Lernende Systeme, München 2021

[274] Herrmann, Thomas, Socio-technical design of hybrid Intelligence systems- the case of predictive maintenance, 2020, avail. at:https://link.springer.com/chapter/10.1007/978-3-030-50334-5_20 (last accessed: 2022-09-26)

[275] Endsley, Mica R., From Here to Autonomy: Lessons Learned From Human–Automation Research, 2016, avail. at: https://journals.sagepub.com/doi/abs/10.1177/0018720816681350 (last accessed: 2022-09-26)

[276] Shneiderman, Ben, Human-Centered Artificial Intelligence: Three Fresh Ideas, 2020, avail. at: https://aisel.aisnet.org/thci/vol12/iss3/1/ (last accessed: 2022-09-26)

[277] Legg, Phil; Smith, Jim; Downing, Alexander, Visual analytics for collaborative human-machine confidence in human-centric active learning tasks, 2019, avail. at: https://hcis-journal.springeropen.com/articles/10.1186/s13673-019-0167-8 (last accessed: 2022-09-26)

[278] Verband der Chemischen Industrie e. V., Industrieland Deutschland, 2022, avail. at: https://www.vci.de/ergaenzende-downloads/industrieland-deutschland-daten-fakten-bedeutung-deutsche-industrie.pdf (last accessed: 2022-08)

[279] VDI/VDE-Gesellschaft Mess- und Automatisierungstechnik, VDI-Statusreport Industrie 4.0 Wertschöpfungsketten, 2014, avail. at: https://www.vdi.de/ueber-uns/presse/publikationen/details/industrie-40-wertschoepfungsketten (last accessed: 2022-09-26)

[280] Positionspapier; Plattform Industrie 4.0, Der Datenraum Industrie 4.0: Die Plattform Industrie 4.0 lädt ein, die digitalen Ökosysteme von morgen zu gestalten, 2021, avail. at: https://www.plattform-i40.de/IP/Redaktion/DE/Downloads/Publikation/PositionPaper-DataSpace.pdf?__blob=publicationFile&v=7 (last accessed: 2022-08)

[281] Anderl, R.; Bauer, K.; Bauernhansel, T.; Diegner, B.; Diemer, J.;Fay, F.; Goericke, D.; Grotepass, J.; Hilger, C.; Jasperneite, J.; Kalhoff, J.; Jubach, U.; Löwen, U.; Menges, G.; Michels, J.S.; Schmidt, F.; Stiedl, T.; ten Hompel, M.; Zeidler, C., Fortschreibung der Anwendungsszenarien der Plattform Industrie 4.0, November 2016, avail. at: https://www.plattform-i40.de/IP/Redaktion/DE/Downloads/Publikation/fortschreibung-anwendungsszenarien.html (last accessed: 2022-09-26)

[282] Bundesministerium für Wirtschaft und Technologie (BMWi), Technologieszenario „Künstliche Intelligenz in der Industrie 4.0, 2019, avail. at: https://www.plattform-i40.de/IP/Redaktion/DE/Downloads/Publikation/KI-industrie-40.html (last accessed: 2022-08)

[283] Geschäftsstelle Plattform Industrie 4.0, Multilaterales Datenteilen in der Industrie: Zielbild am Beispiel des „Collaborative Condition Monitorings" als Basis für neue Geschäftsmodelle, 2022, avail. at: https://www.plattform-i40.de/IP/Redaktion/DE/Downloads/Publikation/Multilaterales_Datenteilen.pdf?__blob=publicationFile&v=8 (last accessed: 2022-09-26)

[284] Anderl, R.; Bauernhansel, T.; Broy, M.; Bullinger-Hoffmann, A.; Eckert, C.; Fay, A.; Gausemeier, J.;Hirsch-Kreinsen, H.; Hornung, G.; Lanza, G.; Liggesmeyer, P.; Nebel, W.; Pfeiffer, S.; Piller, F.; Schildhauer, T.; ten Hompel, M.; Wahlster, W.; Bauer, K.; Bauer, W.; Bond, J., Creutz, S.-M..; Fabian, J.-H., Fehring, A.; Frank, U.; Goericke, D.; Hamann, S.; Kubach, W.; Post, P.; Schöning, H.; von Wichert, G., Themenfelder Industrie 4.0: Forschungs- und Entwicklungsbedarfe zur erfolgreichen Umsetzung von Industrie 4.0, 2019, avail. at: https://www.plattform-i40.de/IP/Redaktion/DE/Downloads/Publikation/acatech-themenfelder-industrie-4-0.pdf?__blob=publicationFile&v=12 (last accessed: 2022-08)

[285] acatech HORIZONTE, KI in der Industrie, 2020, https://www.acatech.de/publikation/acatech-horizonte-ki-in-der-industrie/, (last accessed: 2022-09-26)

[286] Geschäftsstelle Plattform Industrie 4.0, Mit Normen und Standards Industrie 4.0 gestalten, Mai 2022, avail. at: https://www.plattform-i40.de/IP/Redaktion/DE/Downloads/Publikation/Normen-und-Standards.pdf?__blob=publicationFile&v=4 (last accessed: 2022-08)

[287] Geschäftsstelle Plattform Industrie 4.0, Industrie 4.0 gestalten. Resilient, nachhaltig, wettbewerbsstark, Mai 2022, avail. at: https://www.plattform-i40.de/IP/Redaktion/DE/Downloads/Publikation/2022-fortschrittsbericht.pdf?__blob=publicationFile&v=14 (last accessed: 2022-08)

[288] Bundesministerium für Wirtschaft und Technologie (BMWi), KI in der Industrie 4.0: Orientierung, Anwendungsbeispiele, Handlungsempfehlungen, 2019, avail. at: https://www.plattform-i40.de/IP/Redaktion/DE/Downloads/Publikation/ki-in-der-industrie-4-0-orientierung-anwendungsbeispiele-handlungsempfehlungen.html (last accessed: 2022-08)

[289] Jay Lee, Industrial AI: Applications with Sustainable Performance, 2020

[290] Big Data Value Association, Franco-German Position Paper on "Speeding up Industrial AI and Trustworthiness", 2021, avail. at: https://www.bdva.eu/speeding-industrial-ai-and-trustworthiness-position-paper-0 (last accessed: 2022-09)

[291] DIN and DKE, German Standardization Roadmap Industrie 4.0, March 2020, https://www.din.de/resource/blob/65354/1bed7e8d800cd4712d7d1786584a7a3a/roadmap-i4-0-e-data.pdf (last accessed: 2022-09-26)

[292] DIN and DKE, German Standardization Roadmap Industrie 4.0 – Progress Report, April 2022, avail. at: https://www.din.de/resource/blob/868858/73b761333882ebf55ec9aa1ea88fcf43/nrm-industrie-4-0-fortschrittsbericht-en-web-data.pdf (last accessed: 2022-09-26)

[293] ISO/IEC TR 24030:2021, Information technology – Artificial intelligence (AI) – Use cases

[294] PD IEC TR 63283-2:2022, Industrial-process measurement, control and automation. Smart manufacturing. Use cases, avail. at: https://www.vde-verlag.de/iec-normen/250750/iec-tr-63283-2-2022.html (last accessed: 2022-09-26)

[295] Plattform Industrie 4.0, Asset Administration Shell Reading Guide, January 2022, avail. at: https://www.plattform-i40.de/IP/Redaktion/DE/Downloads/Publikation/AAS-ReadingGuide_202201.pdf?__blob=publicationFile&v=4 (last accessed: 2022-09-26)

[296] Bundesministerium für Wirtschaft und Klimaschutz (BMWK), Plattform Industrie 4.0, Details of the Asset Administration Shell: Part 1 – The exchange of information between partners in the value chain of Industrie 4.0 (Version 3.0RC02), May 2022, avail. at: https://www.plattform-i40.de/IP/Redaktion/DE/Downloads/Publikation/Details_of_the_Asset_Administration_Shell_Part1_V3.html (last accessed: 2022-08)

[297] ZVEI, AI in Industrial Automation, April 2021, avail. at: https://www.zvei.org/fileadmin/user_upload/Presse_und_Medien/Publikationen/2021/April/AI_in_Industrial_Automation/AI-in-Industrial-Automation-White-Paper-NEU.pdf (last accessed: 2022-09-26)

[298] Hyde M. Merril et al. IEEE Power & Energy Magazine, PEM pgs. 64 75., Nipping Black outs in the Bud – Introducing a Novel Cascading Failure Network, August 2020, avail. at: https://ieeexplore.ieee.org/document/9120298 (last accessed: 2022-09-26)

[299] Rodrigo Moreno et al. IEEE Power & Energy Magazine, PEM, From Reliability to Resilience – Planning the Grid against the Extremes, August 2020, avail. at: https://ieeexplore.ieee.org/document/9120304 (last accessed: 2022-09-26)

[300] Pfrommer, Julius, Usländer, Thomas und Beyerer, Jürgen, KI-Engineering – AI Systems Engineering: Systematic development of AI as part of systems that master complex tasks" at – Automatisierungstechnik, vol. 70, no. 9, 2022, pp. 756-766, 2022, avail. at: https://doi.org/10.1515/auto-2022-0076 (last accessed: 202209)

[301] Patrik Haslum Australian National University, Nir Lipovetzky University of Melbourne, Daniele Magazzeni King's College London, Christian Muise IBM Research, An Introduction to the Planning Domain Definition Language – Synthesis Lectures on Artificial Intelligence and Machine Learning, 2019, avail. at: https://www.morganclaypool.com/doi/abs/10.2200/S00900ED2V01Y201902AIM042 (last accessed: 2022-09-26)

[302] International Electrotechnical Commission, Semantic interoperability: challenges in the digital transformation age, 2019, avail. at: https://www.iec.ch/basecamp/semantic-interoperability-challenges-digital-transformation-age (last accessed: 202209)

[303] Scientific Data 3, The FAIR Guiding Principles for scientific data management and stewardship, December 2016, avail. at: http://www.nature.com/articles/sdata201618 (last accessed: 2022-08)

[304] AI for Europe – An AI Strategy for Europe, COM(2018)237final, 2018, avail at.: https://eur-lex.europa.eu/legal-content/EN-DE/TXT/?from=de&uri=CELEX%3A52018DC0237 (last accessed: 2022-09-26)

[305] KI-Fachkonferenz von DIN, DKE, Bitkom, VDMA und ZVEI Austausch zum AI Act, 22.11.2021, Fachkonferenz zum Austausch über den AI Act, avail. at: https://www.din.de/de/din-und-seine-partner/presse/mitteilungen/ki-fachkonferenz-von-din-dke-bitkom-vdma-und-zvei–826868 (last accessed: 20220630)

[306] Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on European data governance (Data Governance Act) (Text with EEA relevance)

[307] EU COM(2020)825final – DSA – Ensuring a safe and accountable online environment, EU Digital Services Act, 2020

[308] ISO/IEC SC41/WG6 IoT and Digital Twin, WG6 N089 2nd PWI on Guidelines for IoT and Digital Twin Use Cases

[309] VDE DKE/AK931.0.14_2022-003, SemNorm – Ergebnisse des DINCONNECT Projekts Semantische Normen, 2021

[310] DIN IEC 63351:2022-07 – Draft, VDE 0491-61:2022-07, Nuclear Facilities – Human Factors Engineering – Application to the Design of Human Machine Interfaces

[311] Dede, G., Hamon, R., Junklewitz, H., Naydenov, R., Malatras, A. and Sanchez Martin, J.I., Cybersecurity challenges in the uptake of Artificial Intelligence in Autonomous Driving, EUR 30568 EN, 2021

[312] Christian Berghoff, Jona Böddinghaus, Vasilios Danos, Gabrielle Davelaar, Thomas Doms, Heiko Ehrich, Alexandru Forrai, Radu Grosu, Ronan Hamon, Henrik Junklewitz, Matthias Neu, Simon Romanski, Wojciech Samek, Dirk Schlesinger, Jan-Eve Stavesand, Sebastian Steinbach, Arndt von Twickel, Robert Walter, Johannes Weissenböck, Markus Wenzel and Thomas Wiegand, Towards Auditable AI Systems – From Principles to Practice, Whitepaper, 2022, avail. at: https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/KI/Towards_Auditable_AI_Systems_2022.html (last accessed: 2022-09-26)

[313] Jeannette M. Wing, Trustworthy AI, 2021, Communications of the ACM, Vol. 64 No. 10, Pages 64-71, 10.1145/3448248

[314] Li, Bo and Qi, Peng and Liu, Bo and Di, Shuai and Liu, Jingen and Pei, Jiquan and Yi, Jinfeng and Zhou, Bowen, Trustworthy AI: From Principles to Practices, 2021, arXiv:2110.01167v2, avail. at: https://arxiv.org/abs/2110.01167v2 (last accessed: 2022-08-15)

[315] Davinder Kaur, Suleyman Uslu, Kaley J. Rittichier, and Arjan Durresi, Trustworthy Artificial Intelligence: A Review, 2023, ACM Comput. Surv. 55, 2, Article 39, avail. at: https://doi.org/10.1145/3491209 (last accessed: 2022-08-15)

[316] Fredrik Heintz, Michela Milano and Barry O'Sullivan, Trustworthy AI – Integrating Learning, Optimization and Reasoning, 2021, Lecture Notes in Computer Science, Springer Nature Switzerland AG, avail. at: https://doi.org/10.1007/978-3-030-73959-1 (last accessed: 2022-08-15)

[317] Stanton, B. and Jensen, T., Trust and Artificial Intelligence, 2021, NIST Interagency/Internal Report (NISTIR), National Institute of Standards and Technology, Gaithersburg, MD, avail. at: https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=931087 (last accessed: 2022-08-15)

[318] Christian Berghoff, Battista Biggio, Elisa Brummel, Vasilios Danos, Thomas Doms, Heiko Ehrich, Thorsten Gantevoort, Barbara Hammer, Joachim Iden, Sven Jacob, Heidy Khlaaf, Lars Komrowski, Robert Kröwing, Jan Hendrik Metzen, Matthias Neu, Fabian Petsch, Maximilian Poretschkin, Wojciech Samek, Hendrik Schäbe, Arndt von Twickel, Martin Vechev and Thomas Wiegand, Towards Auditable AI Systems – Current status and future directions, Whitepaper, 2021, Bonn, Berlin, avail. at: https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/KI/Towards_Auditable_AI_Systems.pdf (last accessed: 2022-08-15)

[319] Bundesamt für Sicherheit in der Informationstechnik (BSI), Bundesministerium für Digitales und Verkehr (BMDV), Kraftfahrt-Bundesamt (KBA) und Bundesanstalt für Straßenwesen (BASt), AI-relevant use cases in automotive engineering, 2022, UNECE, 2nd GRVA Workshop on AI and Vehicle Regulations, May 9th 2022, avail. at: https://unece.org/sites/default/files/2022-05/AI-relevant%20use%20cases%20in%20automotive%20engineering.pdf (last accessed: 2022-08-21)

[320] Berghoff C, Neu M and von Twickel A, Vulnerabilities of Connectionist AI Applications: Evaluation and Defense, 2020, avail. at: https://doi.org/10.3389/fdata.2020.00023 (last accessed: 2022-09-26)

[321] Yuan Yang, James C Kerce and Faramarz Fekri, LOGICDEF: An Interpretable Defense Framework Against Adversarial Examples via Inductive Scene Graph Reasoning, 2022

[322] Bundesgesetzblatt, Gesetz zur Änderung des Straßenverkehrsgesetzes und des Pflichtversicherungsgesetzes – Gesetz zum Autonomen Fahren, July 2021, avail. at: https://www.bgbl.de/xaver/bgbl/start.xav?startbk=Bundesanzeiger_BGBl&jumpTo=bgbl121s3108.pdf#__bgbl__%2F%2F*%5B%40attr_id%3D%27bgbl121s3108.pdf%27%5D__1661512907226 (last accessed: 2022-09-26)

[323] Bundesgesetzblatt, Autonome-Fahrzeuge-Genehmigungs-und-Betriebs-Verordnung, June 2022, avail. at: https://www.bgbl.de/xaver/bgbl/start.xav#__bgbl__%2F%2F*%5B%40attr_id%3D%27bgbl122s0986.pdf%27%5D__1661513096702 (last accessed: 2022-09-26)

[324] ISO/SAE 21434:2021, Road vehicles – Cybersecurity engineering

[325] ISO/TR 4804:2020, Road vehicles – Safety and cybersecurity for automated driving systems – Design, verification and validation

[326] ISO/AWI TS 5083, Road vehicles – Safety for automated driving systems – Design, verification and validation

[327] ISO 22737:2021, Intelligent transport systems – Low-speed automated driving (LSAD) systems for predefined routes – Performance requirements, system requirements and performance test procedures

[328] European Union Aviation Safety Agency (EASA), ed., Artificial Intelligence Roadmap 1.0: A human-centric approach to AI in aviation, 2020

[329] European Union Aviation Safety Agency, Concepts of Design Assurance for Neural Networks, March 2020

[330] European Union Aviation Safety Agency, Concepts of Design Assurance for Neural Networks (CoDANN) II, May 2020

[331] Bürkle, A., Segor, F. & Kollmann, M. J Intell Robot Syst 61, 339–353, Towards Autonomous Micro UAV Swarms, 2011, avail. at: https://doi.org/10.1007/s10846-010-9492-x (last accessed: 2022-09-26)

[332] DIN EN 62267:2010-07, VDE 0831-267:2010-07, Railway applications – Automated Urban Guided Transport (AUGT) – Safety requirements (IEC 62267:2009)

[333] Dr. Rainer Müller, Automatische U-Bahn für Nürnberg: Voraussetzungen und Realisierungskonzept der VAG, 1999

[334] Railway Gazette International, Rio Tinto to test Rail Vision collision-avoidance technology, 2021

[335] Ristić-Durrant, Danijela, Marten Franke, and Kai Michels. Sensors 21.10 (2021): 3452, A review of vision-based on-board obstacle detection and distance estimation in railways, 2021

[336] ETSI DGS SAI 003, Securing Artificial Intelligence (SAI); Security Testing of AI

[337] Yvonne Prieur, Andreas Sesing-Wagenpfeil, Christian Müller, Datenschutz beim hochautomatisierten Fahren der Zukunft, in: Tagungsband des Internationalen Rechtsinformatik-Symposiums IRIS 2022

[338] Rustam Tagiew, Thomas Buder, Kai Hofmann, Christian Klotz; eb Ausgabe 6-7, „Risikoanalyse der Schnellbremsung bei frontaler Kollisionsgefahr", 2022

[339] Railway Safety Statistics in the EU, 2022, avail. at: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Railway_safety_statistics_in_the_EU (last accessed: 20220823)

[340] Common Safety Methods – Risk Assessment oder auch für die europäische Durchführungsverordnung Nr. 402/2013 für Allgemeine Sicherheitsmethoden und Risikobewertung, 2013

[341] Bundesministerium für Bildung und Forschung, Ein Multisensorsystem für die Hinderniserkennung Fahrweg: Abschlussbericht; Forschungsvorhaben Komponenten Automatisierter Schienenverkehr (KOMPAS), Phase 1; Arbeitspaket AP 320 Entwicklung Hinderniserkennung, 2003

[342] SAE:J3016:2021, SAE International, Surface Vehicle Recommended Practice – Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles

[343] DIN VDE V 0831-103:2020-09, Electric signalling systems for railways – Part 103: Identification of safety requirements for technical functions in railway signalling

[344] DIN VDE V 0831-101:2022-08, Electric signalling systems for railways – Part 101: Semi-quantitative processes for risk analysis of technical functions in railway signalling

[345] Plattform Lernende Systeme, Kompetenzentwicklung für Künstliche Intelligenz – Veränderungen, Bedarfe und Handlungsoptionen, 2021, avail. at: https://www.acatech.de/publikation/kompetenzentwicklung-fuer-ki-veraenderungen-bedarfe-und-handlungsoptionen/ (last accessed: 2022-09-26)

[346] Europäische Kommission – COM(2021) 205 final, Annexes to the Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions – Fostering a European approach to Artificial Intelligence, 2021, avail. at: https://ec.europa.eu/newsroom/dae/redirection/document/75787 (last accessed: 20220817)

[347] Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S., Dermatologist-level classification of skin cancer with deep neural networks, 2017, avail. at: https://pubmed.ncbi.nlm.nih.gov/28117445/ (last accessed: 2022-09-26)

[348] Muehlematter, UJ, Daniore, P, Vokinger, KN, Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–20): a comparative analysis. The Lancet Digital Health 2021; 3: e195–e203, 2021, avail. at: https://www.thelancet.com/journals/landig/article/PIIS2589-7500(20)30292-2 (last accessed: 2022-09-26)

[349] U.S. Food & Drug Administration, Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices, 2022, avail. at: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices (last accessed: 20220817)

[350] REGULATION (EU) 2017/745 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC (Text with EEA relevance), avail. at: https://eur-lex.europa.eu/legal-content/EN-DE/TXT/?from=DE&uri=CELEX%3A32017R0745 (last accessed: 2022-09-26)

[351] DIN EN ISO 14971:2022, Medical devices – Application of risk management to medical devices (ISO 14971:2019)

[352] ISO/TR 24971:2020, Medical devices – Guidance on the application of ISO 14971

[353] DIN EN 62304:2018, Draft, Health software – Software life cycle processes (IEC 62A/1235/CDV:2018)

[354] DIN EN 82304-1:2018, Health Software – Part 1: General requirements for product safety (IEC 82304-1:2016)

[355] DIN EN 62366-1:2021, Medical devices – Part 1: Application of usability engineering to medical devices (IEC 62366-1:2015 + COR1:2016 + A1:2020)

[356] IG-NB, Leitfaden Künstliche Intelligenz, Version 03.12.2021, Fragenkatalog “Künstliche Intelligenz bei Medizinprodukten“, 2021

[357] ABNT IEC/TR 62366-2:2021, Medical devices – Part 2: Guidance on the application of usability engineering to medical devices

[358] Interessengemeinschaft der Benannten Stellen für Medizinprodukte in Deutschland (IG-NB), Questionnaire “Artificial Intelligence (AI) in medical devices“ (Version 4, 09.06.2022), avail. at: https://www.ig-nb.de/index.php?eID=dump-File&t=f&f=2618&token=010db38d577b0bfa3c909d6f1d74b19485e86975 (last accessed: 20220817)

[359] Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on the European Health Data Space (Text with EEA relevance), 2022, avail. at: https://health.ec.europa.eu/ehealth-digital-health-and-care/europe-an-health-data-space_en (last accessed: 20220817)

[360] DIN EN ISO 14155:2021, Clinical investigation of medical devices for human subjects – Good clinical practice (ISO 14155:2020)

[361] Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, Topol EJ, Ioannidis JPA, Collins GS, Maruthappu M., Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies, 2020, avail. at: BMJ. 2020 Mar 25;368:m689, https://www.bmj.com/content/368/bmj.m689 (last accessed: 2022-09-26)

[362] Cruz Rivera S, Liu X, Chan AW, Denniston AK, Calvert MJ; SPIRIT-AI and CONSORT-AI Working Group; SPIRIT-AI and CONSORT-AI Steering Group; SPIRIT-AI and CONSORT-AI Consensus Group, Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension, 2020, avail at: https://pubmed.ncbi.nlm.nih.gov/32908284/ (last accessed: 2022-09-26)

[363] Liu X, Cruz Rivera S, Moher D, Calvert M, Denniston A, The SPIRIT-AI and CONSORT-AI Working Group, Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension, 2020, avail. at: https://www.nature.com/articles/s41591-020-1034-x (last accessed: 2022-09-26)

[364] Liu X, Glocker B, McCradden MM, Ghassemi M, Denniston AK, Oakden-Rayner L., The medical algorithmic audit, 2022, avail. at: https://www.thelancet.com/journals/landig/article/PIIS2589-7500(22)00003-6/fulltext (last accessed 2022-09-26)

[365] Falk Schwendicke, Tatiana Golla, Martin Dreher, Joachim Krois, Convolutional neural networks for dental image diagnostics: A scoping review, 2019, avail. at: J Dent. 2019 Dec;91:103226. Doi: https://doi.org/10.1016/j.jdent.2019.103226 (last accessed: 20220819)

[366] Taleb A, Rohrer C, Bergner B, De Leon G, Rodrigues JA, Schwendicke F, Lippert C, Krois J., Self-Supervised Learning Methods for Label-Efficient Dental Caries Classification, 2022, avail. at: Diagnostics. 2022; 12(5):1237. https://doi.org/10.3390/diagnostics12051237 (last accessed: 20220819)

[367] M. Wenzel and T. Wiegand, Toward Global Validation Standards for Health AI, 2020, avail. at: IEEE Communications Standards Magazine, vol. 4, no. 3, pp. 64-69, September 2020; doi 10.1109/MCOMSTD.001.2000006 (last accessed: 2022-09-26)

[368] Dudgeon SN, Wen S, Hanna MG, Gupta R, Amgad M, Sheth M, Marble H, Huang R, Herrmann MD, Szu CH, Tong D, Werness B, Szu E, Larsimont D, Madabhushi A, Hytopoulos E, Chen W, Singh R, Hart SN, Sharma A, Saltz J, Salgado R, Gallas BD, A Pathologist-Annotated Dataset for Validating Artificial Intelligence: A Project Description and Pilot Study, 2021, verfügbar unter: J Pathol Inform. 2021 Nov 15;12:45; doi: 10.4103/jpi.jpi_83_20 (last accessed: 2022-09-26)

[369] Guillaume Chassagnon, Maria Vakalopoulou, Enzo Battistella, Stergios Christodoulidis, Trieu-Nghi Hoang-Thi, Severine Dangeard, Eric Deutsch, Fabrice Andre, Enora Guillo, Nara Halm, Stefany El Hajj, Florian Bompard, Sophie Neveu, Chahinez Hani, Ines Saab, Aliénor Campredon, Hasmik Koulakian, Souhail Bennani, Gael Freche, Maxime Barat, Aurelien Lombard, Laure Fournier, Hippolyte Monnier, Téodor Grand, Jules Gregory, Yann Nguyen, Antoine Khalil, Elyas Mahdjoub, Pierre-Yves Brillet, Stéphane Tran Ba, Valérie Bousson, Ahmed Mekki, Robert-Yves Carlier, Marie-Pierre Revel, Nikos Paragios, AI-driven quantification, staging and outcome prediction of COVID-19 pneumonia, 2021, verfügbar unter: Medical Image Analysis, Volume 67, 2021, 101860, ISSN 1361-8415, https://doi.org/10.1016/j.media.2020.101860. (last accessed: 2022-09-26)

[370] Roman Rischke, Lisa Schneider, Karsten Müller, Wojciech Samek, Falk Schwendicke, Joachim Krois, Federated Learning in Dentistry: Chances and Challenges, 2022, avail. at: https://doi.org/10.1177/00220345221108953 (last accessed 2022-09-26)

[371] Ruiyang Ren, Haozhe Luo, Chongying Su, Yang Yao, Wen Liao, Machine learning in dental, oral and craniofacial imaging: a review of recent progress, 2021, avail. at PeerJ 9:e11451 https://doi.org/10.7717/peerj.11451 (last accessed: 20220819)

[372] Jose E Cejudo, Akhilanand Chaurasia, Ben Feldberg, Joachim Krois, Falk Schwendicke, Classification of Dental Radiographs Using Deep Learning, 2021, avail. at: https://doi.org/10.3390/jcm10071496 (last accessed: 2022-09-26)

[373] PD IEC/TR 60601-4-1:2017, Medical electrical equipment – Part 4-1: Guidance and interpretation – Medical electrical equipment and medical electrical systems employing a degree of autonomy

[374] M. Haimerl, Validation of Continuously Learning AI/ML Systems in Medical Devices – A Scenario-based Analysis. Upper Rhine Artificial Intelligence (URAI), 2020

[375] DIN EN 60601-1-10:2021, Medical electrical equipment – Part 1-10: General requirements for basic safety and essential performance – Collateral Standard: Requirements for the development of physiologic closed-loop controllers (IEC 60601-1-10:2007 + A1:2013 + A2:2020)

[376] A.R. Choudhury, R. Vanguri, S.R. Jambawalikar and P. Kumar, Segmentation of Brain Tumors Using DeepLabv3+, 2019, avail. at: https://link.springer.com/chapter/10.1007/978-3-030-11726-9_14 (last accessed: 20220813)

[377] S. Raina, A. Khandelwal, S. Gupta and A. Leekha, Brain Tumor Segmentation Using Unet, 2021, avail. at: https://link.springer.com/chapter/10.1007/978-981-16-1480-4_39 (last accessed: 20220813)

[378] G. Neelima, D. R. Chigurukota, B. Maram, B. Giriajan, Optimal DeepMRSeg based tumor segmentation with GAN for brain tumor classification, 2022, avail. at: https://www.sciencedirect.com/science/article/abs/pii/S1746809422000593 (last accessed: 2022-09-26)

[379] Cohen, J., Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit, 1968, avail. at: https://doi.org/10.1037/h0026256 (last accessed: 2022-09-26)

[380] Landis, J.R., Koch G.G., The Measurement of Observer Agreement for Categorical Data, 1977, avail. at: https://doi.org/10.3389/fnins.2012.00171 (last accessed: 2022-09-26)

[381] DIN EN ISO 13485:2021, Medical devices – Quality management systems – Requirements for regulatory purposes (ISO 13485:2016)

[382] Leonie Beining, KI in der Industrie absichern & prüfen. Was leisten Assurance Cases?, 2021, avail. at: https://www.stiftung-nv.de/sites/default/files/ki_in_der_industrie_sichern_und_prufen.pdf (last accessed: 2022-09-26)

[383] Habli I., Alexander R., Hawkins R. D., Safety Cases: An Impending Crisis?, 2021, avail. at: https://eprints.whiterose.ac.uk/169183/ (last accessed: 2022-09-26)

[384] Marhavilas & Koulouriotis, Risk-Acceptance Criteria in Occupational Health and Safety Risk-Assessment—The State-of-the-Art through a Systematic Literature Review, 2021, avail. at: https://www.mdpi.com/2313-576X/7/4/77/htm (last accessed: 2022-09-26)

[385] Kläs M., Adler R., Jöckel L., Groß J., Reich J., Using Complementary Risk Acceptance Criteria to Structure Assurance Cases for Safety-Critical AI Components, 2021, avail. at: http://ceur-ws.org/Vol-2916/paper_9.pdf (last accessed: 2022-09-26)

[386] European commission, Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on digital operational resilience for the financial sector and amending Regulations (EC) No 1060/2009, (EU) No 648/2012, (EU) No 600/2014 and (EU) No 909/2014 COM/2020/595 final, 2020, avail. at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020PC0595 (last accessed: 2022-09-26)

[387] Bundesministerium für Justiz, Gesetz über Ordnungswidrigkeiten (OwiG) § 111 Falsche Namensangabe, 2021, avail. at: https://www.gesetze-im-internet.de/owig_1968/__111.html (last accessed: 2022-09-21)

[388] CASUALTY ACTUARIAL SOCIETY, CAS RESEARCH PAPER SERIES ON RACE AND INSURANCE PRICING UNDERSTANDING POTENTIAL INFLUENCES OF RACIAL BIAS ON P&C INSURANCE: FOUR RATING FACTORS EXPLORED Members of the 2021 CAS Race and Insurance Research Task Force, 2022, https://www.casact.org/sites/default/files/2022-03/Research-Paper_Understanding_Potential_Influences.pdf?utm_source=Website&utm_medium=Press+Release&utm_campaign=RIP+Series (last accessed: 2022-09-26)

[389] Xi Xing, Fei Huang, Anti-Discrimination Insurance Pricing: Regulations, Fairness Criteria, and Models, 2022, avai. at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3850420 (last accessed: 2022-09-26)

[390] Barocas, S.; Hardt, M.; Narayanan, A., Fairness in Machine Learning. Limitations and Opportunities, 2017, avail. at: Fairness in Machine Learning. Limitations and Opportunities. (last accessed: 2022-09-26)

[391] Hoffmann, Hannah; Vogt, Verena; Hauer, Marc P. et al., Fairness by awareness? On the inclusion of protected features in algorithmic decisions, Preprint 2022, avail. at: Fairness by awareness? (last accessed: 2022-09-26)

[392] European Committee for Standardization (CEN), European Committee for Electrotechnical Standardization (CENELEC), Focus Group Report: Road Map on Artificial Intelligence (AI), 2020, avail. at: https://ftp.cencenelec.eu/EN/EuropeanStandardization/Sectors/AI/CEN-CLC_FGR_RoadMapAI.pdf (last accessed: 20220712)

[393] Dwork, Cynthia; Hardt, Moritz; Pitassi, Tonnian et al., Fairness Through Awareness, 2011, avail. at: Fairness Through Awareness. (last accessed: 2022-09-26)

[394] Kusner M. J., Loftus J. R., Russell C., Silva R., Counterfactual Fairness, 2017, avail. at: https://arxiv.org/abs/1703.06856 (last accessed: 2022-09-26)

[395] Deutsche Bundesbank, Bundesanstalt für Finanzdienstleistungsaufsicht, Maschinelles Lernen in Risikomodellen – Charakteristika und aufsichtliche Schwerpunkte Konsultationspapier, 2021, avail. at: https://www.bundesbank.de/de/startseite/maschinelles-lernen-in-risikomodellen-charakteristika-und-aufsichtliche-schwerpunkte-670944 (last accessed: 2022-09-26)

[396] Kenneth Holmberg & Ali Erdemir, Influence of tribology on global energy consumption, costs and emissions, 2017, avail. at: https://doi.org/10.1007/s40544-017-0183-5 (last accessed: 2022-09-26)

[397] Geibler, Justus von; Gnanko, Toni, Nachhaltige Konsumentscheidungen durch Künstliche Intelligenz und den Digitalen Produktpass – Forschungsbericht zum Roadmapping der Forschungslinie "Transparente Wertschöpfungsketten" im CO:DINA Projekt, 2022, avail. at: https://codina-transformation.de/forschungsbericht_nachhaltige-konsumentscheidungen-durch-kuenstliche-intelligenz-und-den-digitalen-produktpass/ (last accessed: 20220712)

[398] Bundesministerium für Umwelt, Naturschutz, nukleare Sicherheit und Verbraucherschutz, Fünf-Punkte-Programm "Künstliche Intelligenz für Umwelt und Klima", 2021, avail. at: https://www.bmuv.de/download/fuenf-punkte-programm-kuenstliche-intelligenz-fuer-umwelt-und-klima (last accessed: 20220712)

[399] United Nations, Sustainable Development Goals (SDG): 17 Goals to Transform Our World, 2019, avail. at: https://www.un.org/sustainabledevelopment/ (last accessed: 20220712)

[400] Vinuesa, Ricardo; Azizpour, Hossein; Leite, Iolanda; Balaam, Madeline; Dignum, Virginia; Domisch, Sami; Felländer, Anna; Langhans, Simone Daniela; Tegmark, Max; Nerini, Francesco Fuso, The role of artificial intelligence in achieving the Sustainable Development Goals, 2020, avail. at:https://www.nature.com/articles/s41467-019-14108-y (last accessed: 20220712)

[401] Boll, Susanne; Schnell, Markus; Dowling, Michael; Faisst, Wolfgang; Mordvinova, Olga; Pflaum, Alexander; Rabe, Martin; Veith, Eric; Nieße, Astrid; Gülpen, Christian; Terzidis, Orestis; Riss, Uwe, Mit Künstlicher Intelligenz zu nachhaltigen Geschäftsmodellen – Nachhaltigkeit bon, durch und mit KI. Whitepaper aus der Plattform Lernende Systeme, 2022, avail. at: https://www.plattform-lernende-systeme.de/files/Downloads/Publikationen/AG4_WP_KI_und_Nachhaltigkeit.pdf (last accessed: 20220712)

[402] European Commission, European Green Deal, 2019, avail. at: https://ec.europa.eu/info/strategy/priorities-2019-2024/european-green-deal_en (last accessed: 20220712)

[403] Gailhofer, Peter; Herold, Anke; Schemmel, Jan Peter; Scherf, Cara-Sophie; Urrutia, Cristina; Köhler, Andreas R.; Braungardt, Sibylle, The role of Artificial Intelligence in the European Green Deal, 2021, avail. at: https://op.europa.eu/en/publication-detail/-/publication/2c3de271-525a-11ec-91ac-01aa75ed71a1 (last accessed: 20220712)

[404] European Parliament, REGULATION (EU) 2019/2088 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 27 November 2019 on sustainability-related disclosures in the financial services sector

[405] European Parliament, REGULATION (EU) 2020/852 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 18 June 2020 on the establishment of a framework to facilitate sustainable investment, and amending Regulation (EU) 2019/2088

[406] De Lucia, Caterina; Pazienza, Pasquale; Bartlett, Mark, Does Good ESG Lead to Better Financial Performance by Firms? Machine Learning and Logistic Regression Models of Public Enterprises in Europe, 2020

[407] Umweltbundesamt, Jetzke, Tobias; Richter, Stephan; Ferdinand, Jan-Peter; Schaat, Samer, Künstliche Intelligenz im Umweltbereich – Anwendungsbeispiele und Zukunftsperspektiven im Sinne der Nachhaltigkeit, 2019, avail. at: https://www.umweltbundesamt.de/sites/default/files/medien/1410/publikationen/2019-06-04_texte_56-2019_uba_ki_fin.pdf (last accessed: 20220712)

[408] Deutsches Institut für Normung, Deutsche Kommission Elektrotechnik Elektronik, Informationstechnik, Verein Deutscher Ingenieure, Normungslandkarte zur Ressourceneffizienz – Beitrag zu ProgRess III von DIN, DKE und VDI, 2021, avail. at: https://www.din.de/resource/blob/797734/48f084aacb96a7970dd16bfcc88bf53c/normungslandkarte-fuer-ressourceneffizienz-data.pdf (last accessed: 20220712)

[409] Bundesministerium für Umwelt, Naturschutz, nukleare Sicherheit und Verbraucherschutz, Deutsches Ressourceneffizienzprogramm III – 2020 bis 2023, 2020, avail. at: https://www.bmuv.de/fileadmin/Daten_BMU/Pools/Broschueren/ressourceneffizienz_programm_2020_2023.pdf (last accessed: 20220712)

[410] DIN EN ISO 14026:2018, Environmental labels and declarations – Principles, requirements and guidelines for communication of footprint information (ISO 14026:2017)

[411] DIN EN ISO 14040:202102, Environmental management – Life cycle assessment – Principles and framework (ISO 14040:2006 + Amd 1:2020)

[412] DIN EN ISO 14044:2021, Environmental management – Life cycle assessment – Requirements and guidelines (ISO 14044:2006 + Amd 1:2017 + Amd 2:2020)

[413] DIN EN 15804:2022, Sustainability of construction works – Environmental product declarations – Core rules for the product category of construction products

[414] DIN EN ISO 22057:2022, Sustainability in buildings and civil engineering works – Data templates for the use of environmental product declarations (EPDs) for construction products in building information modelling (BIM) (ISO 22057:2022)

[415] DIN EN 62559-2:2016-05; VDE 0175-102:2016-05, Use case methodology – Part 2: Definition of the template for use cases, actor list and requirements list (IEC 62559-2:2015)

[416] Clauß, John; Finck, Christian; Vogler-Finck, Pierre; Beagon, Paul, Control strategies for building energy systems to unlock demand side flexibility – A review, 2017

[417] Bundesministerium für Umwelt, Naturschutz und nukleare Sicherheit (BMU), Nationales Programm für nachhaltigen Konsum. Gesellschaftlicher Wandel durch einen nachhaltigen Lebensstil, 2019, avail. at: https://nachhaltigerkonsum.info/sites/default/files/medien/dokumente/nachhaltiger_konsum_broschuere_bf.pdf (last accessed: 20220811)

[418] Gossen, Maike, Jankowski, Patricia, Driving Sustainable Behavior with Persuasive Technology: The Green Consumption Assistant, 2022. Ökologisches Wirtschaften, 2.2022 (37)

[419] Kahl, G.; Herbig, N.; Erdmann, L.; Stadler, K.; Peters, A., Ergebnisdokumentation des Praxisprojekts “Kundenführung am Point of Sale“: Arbeitspapier im Arbeitspaket 4 (AP 4.4) des INNOLAB Projekts. Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI GmbH), Saarbrücken 2017

[420] Altmeyer, Maximilian; Schubhan, Marc; Kerber, Frederic, Automatisieren Personalisieren, Optimieren: Chancen & Herausforderungen von KI-Anwendungen auf Basis des Digitalen Produktpasses im Handel, 2022, avail. at: https://codina-transformation.de/kurzstudie/ (last accessed: 202208-11)

[421] Bundesministerium für Umwelt, Naturschutz und nukleare Sicherheit (BMU), Digitaler Produktpass, 2021, avail. at: https://www.bmu.de/faqs/umweltpolitische-digitalagenda-digitaler-produktpass/ (last accessed: 20220811)

[422] Götz, Thomas; Berg, Holger; Jansen, Maike; Adisorn, Thomas; Cembrero, David; Markkanen, Sanna; Chowdhury, Tahmid, Digital Product Passport: the ticket to achieving a climate neutral and circular European economy?, 2022, avail. at: https://www.corporateleadersgroup.com/files/cisl_digital_products_passport_report_v6.pdf (last accessed: 20220811)

[423] Geibler, Justus von; Gnanko, Toni, Künstliche Intelligenz für nachhaltigen Konsum. Ansatzpunkte und Herausforderungen für nachhaltige Konsumentscheidungen auf Basis künstlicher Intelligenz, 2021, avail. at: https://codina-transformation.de/wp-content/uploads/CO-DINA-Positionspapier-7-KI-und-Nachhaltiger-Konsum-1.pdf (last accessed: 20220811)

[424] Lasarov, Wassili, Nachhaltiger Konsum im digitalen Zeitalter, 2022, In: Bruhn, M., Hadwich K. (ed.), Künstliche Intelligenz im Dienstleistungsmanagement. Springer Fachmedien Wiesbaden GmbH, 235–262

[425] Schneider-Marin, Patricia; Harter, Hannes; Tkachuk, Konstantin; Lang, Werner, Uncertainty Analysis of Embedded Energy and Greenhouse Gas Emissions Using BIM in Early Design Stages, 2020

[426] ISO/IEC 19763-3:2020, Information technology – Metamodel framework for interoperability (MFI) – Part 3: Metamodel for ontology registration

[427] DIN/TS 92004, Artificial intelligence – Quality requirements and processes – Risk scheme for AI systems along the entire life cycle

[428] ISO/IEC PWI 7699, Guidance for addressing security threats and failures in artificial intelligence

[429] ISO/IEC 2382:2015, Information technology – Vocabularies

[430] DIN EN IEC 81001-5-1:2022-01 – Draft, VDE 0750-103-5-1:2022-01, Health Software and health IT systems safety, effectiveness and security – Part 5-1: Security – Activities in the product lifecycle (IEC 62A/1419/CDV:2020)

[431] ISO/IEC 20924:2021, Internet of things (IoT) – Vocabulary

[432] ISO/IEC AWI 42005, Information technology – Artificial intelligence – AI system impact assessment

[433] DIN EN 61508-5:2011-02, VDE 0803-5:2011-02, Functional safety of electrical/electronic/programmable electronic safety-related systems – Part 5: Examples of methods for the determination of safety integrity levels (IEC 61508-5:2010)

[434] DIN EN 61511-1:2019-02, VDE 0810-1:2019-02, Functional safety – Safety instrumented systems for the process industry sector – Part 1: Framework, definitions, system, hardware and application programming Requirements (IEC 61511-1:2016 + COR1:2016 + A1:2017)

[435] DIN EN IEC 62443 (all parts), Security for industrial automation and control systems

[436] ISO/IEC TR 24027:2021, Information technology – Artificial intelligence (AI) – Bias in AI systems and AI aided decision making

[437] ISO/IEC TR 24372:2021, Information technology – Artificial intelligence (AI) – Overview of computational approaches for AI systems

[438] ISO/IEC TR 20547-1:2020, Information technology – Big data reference architecture – Part 1: Framework and application process

[439] ISO/IEC TR 20547-2:2018, Information technology – Big data reference architecture – Part 2: Use cases and derived requirements

[440] ISO/IEC 20547-3:2020, Information technology – Big data reference architecture – Part 3: Reference architecture

[441] ISO/IEC 20547-4:2020, Information technology – Big data reference architecture – Part 4: Security and privacy

[442] ISO/IEC TR 20547-5:2018, Information technology – Big data reference architecture – Part 5: Standards roadmap

[443] ISO/IEC 20546:2019, Information technology – Big data – Overview and vocabulary

[444] ISO/IEC 33063:2015, Information technology – Process assessment – Process assessment model for software testing

[445] DIN EN ISO/IEC 15408-1:2020, Information technology – Security techniques – Evaluation criteria for IT security – Part 1: Introduction and general model (ISO/IEC 15408-1:2009)

[446] DIN EN ISO/IEC 15408-2:2020, Information technology – Security techniques – Evaluation criteria for IT security – Part 2: Security functional components (ISO/IEC 15408-2:2008)

[447] DIN EN ISO/IEC 15408-3:2020, Information technology – Security techniques – Evaluation criteria for IT security – Part 3: Security assurance components (ISO/IEC 15408-3:2008, Corrected version 2011-06-01)

[448] ISO/IEC 15408-4:2022, Information security, cybersecurity and privacy protection – Evaluation criteria for IT security – Part 4: Framework for the specification of evaluation methods and activities

[449] ISO/IEC 15408-5:2022, Information security, cybersecurity and privacy protection – Evaluation criteria for IT security – Part 5: Pre-defined packages of security requirements

[450] ETSI TR 101 583:2015, Methods for Testing and Specification (MTS) – Security Testing – Basic Terminology

[451] DIN EN 61513:2013-09, VDE 0491-2:2013-09, Nuclear power plants – Instrumentation and control important to safety – General requirements for systems (IEC 61513:2011)

[452] DIN EN 50128:2012-03; VDE 0831-128:2012-03, Railway applications – Communication, signalling and processing systems – Software for railway control and protection systems

[453] IEEE 7010:2020, A New Standard for Assessing the Well-being Implications of Artificial Intelligence

[454] IEEE P2801:2021, Recommended Practice for the Quality Management of Datasets for Medical Artificial Intelligence

[455] ISO 26262 (all parts), Road vehicles – Functional safety

[456] DIN EN 62061:2017, Draft, Safety of machinery – Functional safety of safety-related electrical, electronic and programmable electronic control systems (IEC 44/788/CD:2017)

[457] DIN CEN ISO/TR 22100-1:2021, Safety of machinery – Relationship with ISO 12100 – Part 1: How ISO 12100 relates to type-B and type-C standards (ISO/TR 22100-1:2021)

[458] DIN ISO/TR 22100-2, DIN SPEC 33887, Safety of machinery – Relationship with ISO 12100 – Part 2: How ISO 12100 relates to ISO 13849-1

[459] DIN ISO/TR 22100-3, DIN SPEC 33888, Safety of machinery – Relationship with ISO 12100 – Part 3: Implementation of ergonomic principles in safety standards

[460] DIN CEN ISO/TR 22100-4:2020, Safety of machinery – Relationship with ISO 12100 – Part 4: Guidance to machinery manufacturers for consideration of related IT-security (cyber security) aspects (ISO/TR 22100-4:2018)

[461] ISO/TR 22100-5:2021, Safety of machinery – Relationship with ISO 12100 – Part 5: Implications of artificial intelligence machine learning

[462] DIN EN ISO 13849-2:2013, Safety of machinery – Safety-related parts of control systems – Part 2: Validation (ISO 13849-2:2012)

[463] ISO/IEC 25012:2008, Software engineering – Software product Quality Requirements and Evaluation (SQuaRE) – Data quality model

[464] ISO/IEC/IEEE 29119-1:2022, Software and systems engineering – Software testing – Part 1: General concepts

[465] ISO/IEC/IEEE 29119-2:2021, Software and systems engineering – Software testing – Part 2: Test processes

[466] ISO/IEC/IEEE 29119-3:2021, Software and systems engineering – Software testing – Part 3: Test documentation

[467] ISO/IEC/IEEE 29119-4:2021, Software and systems engineering – Software testing – Part 4: Test techniques

[468] ISO/IEC/IEEE 29119-5:2016, Software and systems engineering – Software testing – Part 5: Keyword-driven testing

[469] IEEE 1012:2016, Standard for System, Software, and Hardware Verification and Validation

[470] IEEE 3333.1.3:2022, Standard for the Deep Learning-Based Assessment of Visual Experience Based on Human Factors

[471] ANSI/UL 4600:2022, Standard for Safety for the Evaluation of Autonomous Products

[472] ISO/IEC 25000:2014, Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – Guide to SQuaRE

[473] ISO/IEC 25024:2015, Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – Measurement of data quality

[474] ISO/IEC 25020:2019, Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – Quality measurement framework

[475] ISO/IEC 25021:2012, Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – Quality measure elements

[476] DIN SPEC 2343:2020, Transmission of language-based data between artificial intelligences – Specification of parameters and formats

[477] ISO/TS 17033:2019, Ethical claims and supporting information – Principles and requirements

[478] DIN EN ISO 26000:2021, Guidance on social responsibility (ISO 26000:2010)

[479] DIN EN ISO/IEC 27000:2020, Information technology – Security techniques – Information security management systems – Overview and vocabulary

[480] DIN EN ISO/IEC 27001:2017, Information technology – Security techniques – Information security management systems – Requirements

[481] DIN EN ISO/IEC 27002:2022 – Draft, Information security, cybersecurity and privacy protection – Information security controls (ISO/IEC 27002:2022)

[482] ITU-T Y gos-ml-arc, Architecture of machine learning based QoS assurance for the IMT-2020 network, 2019 draft

[483] ETSI TS 103 195-2:2018, Autonomic network engineering for the self-managing Future Internet (AFI); Generic Autonomic Network Architecture; Part 2: An Architectural Reference Model for Autonomic Networking, Cognitive Networking and Self-Management, avail. at: https://portal.etsi.org/webapp/WorkProgram/Report_WorkItem.asp?WKI_ID=50970 (last accessed: 2022-09-26)

[484] DIN EN ISO/IEC 17021-2:2019, Conformity assessment – Requirements for bodies providing audit and certification of management systems – Part 2: Competence requirements for auditing and certification of environmental management systems (ISO/IEC 17021-2:2016)

[485] DIN EN ISO/IEC 17021-3:2019, Conformity assessment – Requirements for bodies providing audit and certification of management systems – Part 3: Competence requirements for auditing and certification of quality management systems (ISO/IEC 17021-3:2017)

[486] DIN EN ISO/IEC 17030:2021, Conformity assessment – General requirements for third-party marks of conformity (ISO/IEC 17030:2021)

[487] DIN EN ISO/IEC 17040:2005, Conformity assessment – General requirements for peer assessment of conformity assessment bodies and accreditation bodies (ISO/IEC 17040:2005)

[488] DIN EN ISO/IEC 17043:2022, Draft, Conformity assessment – General requirements for the competence of proficiency testing providers (ISO/IEC DIS 17043:2022)

[489] DIN EN ISO/IEC 17050-1:2010, Conformity assessment – Supplier's declaration of conformity – Part 1: General requirements

[490] DIN EN ISO/IEC 17050-2:2005, Conformity assessment – Supplier's declaration of conformity – Part 2: Supporting documentation (ISO/IEC 17050-2:2004)

[491] ITU-T F.AI-DLFE, Deep Learning Software Framework Evaluation Methodology, 2021

[492] ITUT Y.3173, Framework for evaluating intelligence level of future networks including IMT-2020, 2020

[493] DIN EN ISO/IEC 29101:2022, Information technology – Security techniques – Privacy architecture framework (ISO/IEC 29101:2018)

[494] DIN EN ISO/IEC 29147:2020, Information technology – Security techniques – Vulnerability disclosure (ISO/IEC 29147:2018)

[495] ITU-T F.AI-DLPB, Metrics and evaluation methods for deep neural network processor benchmark, 2020

[496] ITU-T Y.3170, Requirements for machine learning – based quality of service assurance for the IMT-2020 Network, 2018

[497] ETSI DGR SAI 002, Securing Artificial Intelligence (SAI); Data Supply Chain Report, 2021,

[498] ETSI TS 103 296, Speech and Multimedia Transmission Quality (STQ); Requirements for Emotion Detectors used for Telecommunication Measurement Applications; Detectors for written text and spoken speech, 2016

[499] ETSI GR ENI 004, Experiential Networked Intelligence (ENI); Terminology for Main Concepts in ENI Disclaimer, 2019

[500] ETSI GR NFV 003, Network Functions Virtualisation (NFV); Terminology for Main Concepts in NFV, 2020

[501] ISO/TR 24291:2021, Health informatics – Applications of machine learning technologies in imaging and other medical applications

[502] ISO/TR 3985:2021, Biotechnology – Data publication – Preliminary considerations and concepts

[503] ISO/TS 22756:2020, Health Informatics – Requirements for a knowledge base for clinical decision support systems to be used in medication-related processes

[504] ITU-T F.VS-AIMC, Use cases and requirements for multimedia communication enabled vehicle systems using artificial intelligence, 2021

[505] DIN SPEC 91426:202012, Quality requirements for video-based methods of personnel selection

[506] DIN SPEC 13288:2021, Guideline for the development of deep learning image recognition systems in medicine

[507] ISO/TS 5346:2022, Health informatics – Categorial structure for representation of traditional Chinese medicine clinical decision support system

[508] DIN EN ISO 11073 series, Health informatics – Personal health device communication

[509] DIN CEN ISO/TS 22703:2022, Health informatics – Requirements for medication safety alerts (ISO/TS 22703:2021)

[510] ISO/TR 19669:2017, Health informatics – Re-usable component strategy for use case development

[511] IEEE P2802:2022, Standard for the Performance and Safety Evaluation of Artificial Intelligence Based Medical Device: Terminology

[512] VDI-MT 7001:2021, Communication and public participation in construction and infrastructure projects – Standards for work stages of engineers

[513] DIN EN ISO 10075 (all parts), Ergonomic principles related to mental workload

[514] DIN EN ISO 9241 (all parts), Ergonomic requirements for office work with visual display terminals (VDTs)

[515] DIN EN 894 (all parts), Safety of machinery – Ergonomics requirements for the design of displays and control actuators

[516] DIN EN 16710-2:2016, Ergonomics methods – Part 2: A methodology for work analysis to support design

[517] DIN EN ISO 12100:2011, Safety of machinery – General principles for design – Risk assessment and risk reduction (ISO 12100:2010)

[518] ISO/TR 16982:2002, Ergonomics of human-system interaction – Usability methods supporting human-centred design

[519] ISO 21930:2017, Sustainability in buildings and civil engineering works – Core rules for environmental product declarations of construction products and services

[520] DIN IEC/TS 62998-1:2021-10, VDE V 0113-998-1:2021, Safety of machinery – Safety-related sensors used for the protection of persons (IEC TS 62998-1:2019)

[521] ISO/TS 14048:2002, Environmental Management – Life Cycle Assessment – Data Documentation Format

[522] CWA 17284:2018, Materials modelling – Terminology, classification and metadata

[523] CWA 17815:2021, Materials characterization – Terminology, metadata and classification

[524] Geibler, Justus von; Riera, Nuria; Echternacht, Laura; Björling, Sten-Eric.; Domen, Tom et al., myEcoCost. Forming the Nucleus of a Novel Environmental Accounting System: Vision, prototype and way forward, 2015, Wuppertal Institute for Climate, Environment and Energy, Wuppertal, avail. at: https://epub.wupperinst.org/frontdoor/index/index/docId/6009 (last accessed: 20220712)

11 List of authors

Dr.-Ing. Rasmus Adler, Fraunhofer-Institut für Experimentelles Software Engineering (IESE)

Araceli Alcala, Carl Zeiss Meditec AG

Marie Anton, Bundesverband der Arzneimittel-Hersteller e. V. (BAH)

Tristan Armbruster, PricewaterhouseCoopers GmbH

Stefan Arntzen, Huawei Technologies Düsseldorf GmbH

Prof. Dr. Doris Aschenbrenner, Aalen University

Klaus-Dieter Axt, EUnited (European Engineering Industries Association)

Dr. Renate Baumgartner, Eberhard Karls University Tübingen

Nikolas Becker, Gesellschaft für Informatik e. V. (GI)

Rebecca Beiter, Cyber Valley

Thomas Bendig, adesso SE

Ralf Benecke, sonnen GmbH

Dr. Philipp Benner, Bundesanstalt für Materialforschung und -prüfung (BAM)

Bastian Bernhardt, IABG mbH

Paul Beyer, FSD Fahrzeugsystemdaten GmbH

Karsten Bich, German Institute for Standardization (DIN)

Jan Biehler, Plattform Lernende Systeme / acatech

Lukas Bieringer, QuantPi GmbH

Dr. Andreas Binder, SAMSON Pilotentwicklung GmbH

Dr. Sylwia Birska, BG ETEM

André Bluhm, ai.dopt GmbH

John Böhm, T-Systems Multimedia Solutions GmbH

Dr. Jürgen Bohn, Schaeffler AG

Dr.-Ing. Patrick Bollgrün, Plattform Lernende Systeme / acatech

Dr.-Ing. Mikko Börkircher, Verband der Metall- und ElektroIndustrie Nordrhein-Westfalen e. V. (METALL NRW)

Kevin Borowski, HASPA

Oliver Bracht, eoda GmbH

Matthias Brand, MBDA Deutschland GmbH

Katharina Buchsbaum, German Research Institute for Public Administration (FÖV)

Lena Marie Budde, Bund für Umwelt und Naturschutz Deutschland (BUND)

Dr. Joachim Bühler, TÜV-Verband

Dr. Andreas Bunte, Fraunhofer IOSB-INA (IOSB-INA)

Dr. Aljoscha Burchardt, German Research Center for Artificial Intelligence (DFKI)

Prof. Dr. Simon Burton, Fraunhofer-Institut für Kognitive Systeme (IKS)

Tim Büttel, TÜV Nord Mobilität GmbH & Co. KG IFM

Ulla Coester, xethix Empowerment

Prof. Dr. Armin B. Cremers, b-it Emeritus Research Group AI Foundations, Universität Bonn

Damian A. Czarny, German Commission for Electrical, Electronic & Information Technologies of DIN and VDE (DKE)

Lucas da Silva, Ingrano Solutions GmbH

Dr. David Dang, Deloitte GmbH

Klaus Däßler, Gesellschaft für Mathematische Intelligenz (GMI)

Jan de Meer, University of Applied Sciences Berlin (HTW)

Axel Demel, qdive GmbH

Dr. Peter Deussen, Microsoft Deutschland GmbH

Ernestine Dickhaut, Kassel University

Eckhard Dittrich, private person

Alexander Dittrich, Deloitte GmbH

Lilian Do Khac, Philipps-Universität Marburg und adesso SE

Felix Dotzauer, SPECTARIS – Deutscher Industrieverband für Optik, Photonik, Analysen- und Medizintechnik e. V.

Gilbert Drzyzga, Technische Hochschule Lübeck

Prof. Dr. Martin Ebers, Robotics & AI Law Society (RAILS)

Filiz Elmas, DIN German Institute for Standardization

Dr. Stefan Elmer, Festo SE & Co. KG

Dr.-Ing. Marko Esche, German National Metrology Institute (PTB)

Benjamin Fehlandt, SALT AND PEPPER Technology

Leander Féret, JUMO GmbH & Co. KG

Lajla Fetic, Bertelsmann Stiftung

Marc Fliehe, TÜV-Verband e. V.

Dr. Julia Fligge-Niebling, German Aerospace Center (DLR)

Werner Flögel, GEMÜ Gebr. Müller Apparatebau GmbH & Co. KG

Christopher Frank, Deutsche Gesetzliche Unfallversicherung (DGUV)

Matthias Frank, Brose Fahrzeugteile SE & Co. KG

Annika Franken, Forschungsinstitut für Rationalisierung (FIR) e. V. at RWTH Aachen University

Prof. Dr. Martin Fränzle, Oldenburg University

Christian Fraunholz, Fraunholz Technologies UG

Charlotte Frierson, Forschungsinstitut für Rationalisierung (FIR) e. V. at RWTH Aachen University

Florian Gauer, PricewaterhouseCoopers GmbH

Prof. Dr. Clemens Gause, Verband für Sicherheitstechnik e. V. (VFS)

Antoine Gautier, QuantPi GmbH

Dr. Marc Gebauer, Brandenburg Technical University Cottbus-Senftenberg

Dr. Bernd Geiger, semafora systems GmbH

Dr. Sergio Genovesi, Bonn University

Dr. Detlef Gerst, IG Metall

Simon Geschwill, Schwarz Dienstleistung KG

Prof. Dr. Dagmar Gesmann-Nuissl, University of Technology Chemnitz (TUC)

Nora Helena Glasmeier, Bundesverband der Deutschen Volksbanken und Raiffeisenbanken e. V. (BVR)

Dr. Ludwig Glatzner, Büro für Umwelt, Qualität, Sicherheit

Jens Gnaudschun, TÜV Nord Mobilität GmbH & Co. KG

Marius Goebel, Spherity GmbH

Dominik Grau, Beuth Verlag

Viacheslav Gromov, AITAD GmbH

Dr.-Ing. Jürgen Großmann, Fraunhofer-Institut für Offene Kommunikationssysteme (FOKUS)

Prof. Dr. Jürgen Grotepass, Huawei Technologies Düsseldorf GmbH

Yvonne Gruchmann, Wirtschaftsförderung Land Brandenburg GmbH

Norman Günther, Wildau Technical University of Applied Sciences (TH Wildau)

Prof. Dr. Martin Haimerl, Furtwangen University (HFU)

Christian Hattenkofer, Bank-Verlag GmbH

Marc Hauer, Algorithm Accountability Lab der TU Kaiserslautern (AAL TUK)

Prof. Dr. Stefan Haufe, National Metrology Institute (PTB) and Technical University of Berlin

Elias Heider, MARELLI

Jürgen Heiles, Siemens AG

Dr. Tobias Heimann, Siemens Healthineers

Tabea Hein, Stadtverwaltung Frankfurt am Main

Claudia Heinemann, Independent consultant

Christoph Henseler, Deutsches Institut für Gutes Leben GmbH (difgl)

Dr. Wolfgang Hildesheim, IBM Germany GmbH

Barbara Hilgert, Fortbildungsakademie der Wirtschaft (FAW) gemeinnützige Gesellschaft mbH

Dr. Lukas Höhndorf, IABG mbH

Taras Holoyad, German Federal Network Agency

Dr. Maximilian Hösl, Plattform Lernende Systeme / acatech

Alexander Jaschke, Fraunhofer-Institut für Integrierte Schaltung (IIS)

Dr. Barbara Jung, German National Metrology Institute (PTB)

Dr. Vanessa Just, KI Bundesverband e. V.

Agnieszka Kacyniak, consileo GmbH & Co. KG

Thomas Kaiser, Kodex AI GmbH

Dr. Leo Kärkkäinen, Huawei Technologies Deutschland GmbH

Jan Kiefer, German Federal Financial Supervisory Authority (BaFin)

So-Jin Kim, German Institute for Standardization (DIN)

Roland Kirsch, German Federal Office for Information Security (BSI)

Nils-Olaf Klabunde, 4PL Intermodal GmbH

Dr. Michael Kläs, Fraunhofer-Institut für Experimentelles Software Engineering (IESE)

Philip Kleen, Fraunhofer-Institutsteil für industrielle Automation (INA) des Fraunhofer IOSB

Prof. Dr. Annette Kleinfeld, Hochschule für Technik, Wirtschaft und Gestaltung (HTWG) Konstanz

Anita Klingel, PD – Consultant for the public sector

Dr. habil. Jürgen Klippert, IG Metall

Julia Kloiber, Superrr Lab

Mirko Knaak, IAV GmbH

David Knauer, T-Systems Multimedia Solutions GmbH

Ricardo Knauer, University of Applied Sciences in Berlin (HTW)

Andrea Knaut, Institut für Sozialarbeit und Sozialpädagogik e. V./ Geschäftsstelle Dritter Gleichstellungsbericht der Bundesregierung

Franz Knecht, Connexis AG

Mario Knicker, SHARP Electronics GmbH

Marco Knödler, YNCORIS GmbH & Co. KG / Interessensgemeinschaft Regelwerke Technik e. V. (IGR)

Harry Knopf, High Knowledge GmbH

Dr. Martin F. Köhler, Lawyer

Christian Kolf, TÜV AI Lab

Dr. Georgios Kolliarakis, German Council on Foreign Relations (DGAP)

Christopher Koska, dimension2 economics & philosophy consult GmbH

Sebastian Kosslers, German Commission for Electrical, Electronic & Information Technologies of DIN and VDE (DKE)

Roland Kossow, CyberTribe® – Das dezentrale Systemhaus

Dr. Stefan Kothe, German National Metrology Insitute (PTB)

Sebastian Kotte, neurocat GmbH

Tobias Krafft, Trusted AI GmbH

Prof. Dr.-Ing. Klaus Kratzer, Ulm University of Applied Sciences

Dr. Tom Kraus, VDI/VDE Innovation und Technik GmbH

Prof. Markus Krebsz, UNECE GRM & The Human-Ai.Institute

Tim Kremer, Deutscher Sparkassen und Giroverband e. V.

Sebastian Kriegsmann, German Institute for Standardization (DIN)

Mirco Kröll, Bundesanstalt für Materialforschung und -prüfung (BAM)

Prof. Dr. Antonio Krüger, German Research Centre for Artificial Intelligence (DFKI)

Katja Krüger, German Institute for Standardization (DIN)

Jacques Kruse Brandao, SGS

Dr. Tanja Kubes, Freie Universität Berlin

Susanne Kuch, Deutsche Akkreditierungsstelle GmbH (DAkkS)

Stefan Kunkel, sagena Innovationsgesellschaft mbH

Benjamin Küttner, Deutsche Bundesbank

Dr. Jens F. Lachenmaier, University of Stuttgart

Holger Laible, Siemens AG

Joel Lakermann, TÜV Nord Mobilität GmbH & Co. KG

Fredi Lang, Berufsverband Deutscher Psychologinnen und Psychologen e. V.

Dr. Erich Latniak, Universität Duisburg-Essen

Elisa Lederer, PricewaterhouseCoopers GmbH

Dr.-Ing. Christoph Legat, HEKUMA GmbH

Lorenz Lehmhaus, Aleph Alpha GmbH

Dr. Mahei Manhai Li, Kassel University

Michael Lipka, Huawei Technologies Deutschland GmbH

Daniel Loevenich, German Federal Office for Information Security (BSI)

PD Dr. habil. Jeanette Miriam Lorenz, Fraunhofer-Institut für Kognitive Systeme (IKS)

Prof. Dr. Ulrich Löwen, Siemens AG

Dr. Jackie Ma, Fraunhofer-Institut für Nachrichtentechnik, Heinrich-Hertz-Institut (HHI)

Dr.-Ing. Stefan Maack, Bundesanstalt für Materialforschung und -prüfung (BAM)

Manuela Mackert, private person

Sabine Mahr, word b sign Sabine Mahr

Maximilian Margreiter, Deloitte GmbH

Karla Markert, Fraunhofer-Institut für Angewandte und Integrierte Sicherheit (AISEC)

Dr. Oliver Maspfuhl, Deutsche Bank AG

Isabel Matthias, Bremen University

Gerd Matzke, Drägerwerk AG & Co. KGaA

Henri Meeß, Fraunhofer-Institut für Verkehrs- und Infrastruktursysteme (IVI)

Iris Merget, German Research Centre for Artificial Intelligence (DFKI)

Ralf Meschede, Federal Highway Research Institute (BASt)

Martin Meyer, Siemens Healthineers

Olga Meyer, Fraunhofer-Institut für Produktionstechnik und Automatisierung (IPA)

Dr.-Ing. Sascha Meyne, German National Metrology Institute (PTB)

Alexander Mihatsch, Plattform Lernende Systeme / acatech

Dr. Alexander G. Mirnig, Paris Lodron Universität Salzburg & AIT Austrian Institute of Technology GmbH

Prof. Dr. Andreas Mockenhaupt, Albstadt-Sigmaringen University

Dr.-Ing. Eike Möhlmann, German Aerospace Center (DLR)

Michael Mörike, Integrata-Stiftung

Andreas Müller, Schaeffler AG

Dr. Christian Müller, German Research Center for Artificial Intelligence (DFKI)

Tomislav Nad, SGS

Gert Nahler, Samson AG

Dr. Andreas Nawroth, Münchener Rückversicherungs-Gesellschaft (Munich RE)

Jens Neuhüttler, Fraunhofer-Institut für Arbeitswirtschaft und Organisation (IAO)

Dr. Matthias Neumann-Brosig, IAV GmbH

Dr. Marc Neveling, Deloitte GmbH

Dr. Peter Nickel, Institut für Arbeitsschutz der Deutschen Gesetzlichen Unfallversicherung (IFA)

Jürgen Niehaus, SafeTRANS

Reimund Nienaber, EDLIGO

Johannes Nöbel, KPMG AG Wirtschaftsprüfungsgesellschaft

Dr. Antje Nowack, Verband der Vereine Creditreform e. V.

Dr. Shane O'Sullivan, Universidade de São Paulo

Otto Obert, Main DigitalEthiker GmbH

Dr. Ursula Ohliger, Plattform Lernende Systeme / acatech

Rebecca Page, Endress+Hauser Flowtec AG

Dr. Jochen Papenbrock, NVIDIA GmbH

Ludwig Pechmann, UniTransferKlinik Lübeck GmbH

Yannick Peifer, Institut für angewandte Arbeitswissenschaft e. V. (ifaa)

Robin H. Pekerman, swiss4digital

Dr. Annelie Pentenrieder, Institut für Innovation und Technik Berlin

Dr. Christoph Peters, Kassel University – ITeG

Fabian Petsch, German Federal Office for Information Security (BSI)

Katharina Petschick, WBS GRUPPE

Dr. Christoph Peylo, Robert Bosch GmbH

Daniel Pflumm, TÜV-Verband e. V.

Patrick Philipp, Fraunhofer Institute of Optronics, System Technologies and Image Exploitation (IOSB)

Dr. Henrich C. Pöhls, University of Passau

Frank Poignée, infoteam Software AG

Dr. Maximilian Poretschkin, Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme (IAIS)

Martin Portier, Bundesamt für Seeschifffahrt und Hydrographie (BSH)

Dr.-Ing. Jens Prager, Bundesanstalt für Materialforschung und -prüfung (BAM)

Dr. Henrik J. Putzer, fortiss

Alexander Rabe, eco – Verband der Internetwirtschaft e. V.

Peter Rauh, German Institute for Standardization (DIN)

Hendrik A. Reese, PricewaterhouseCoopers GmbH

Prof. Dr. Georg Rehm, German Research Center for Artificial Intelligence (DFKI)

Dr. Claudia Reinel, German Institute for Standardization (DIN)

Axel Rennoch, Fraunhofer-Institut für Offene Kommunikationssysteme (FOKUS)

Klaus Roleff, Wintegral GmbH

Karsten Roscher, Fraunhofer-Institut für Kognitive Systeme (IKS)

Michael Rosenthal, regio iT gesellschaft für informationstechnologie mbh

Jan Rösler, German Institute for Standardization (DIN)

Nils Röttger, imbus AG

Dr. Gerhard Runze, imbus AG

Martin Ruskowski, DFKI

Dr. Martin Saerbeck, TÜV SÜD

Peter K. Sanner, areasix GmbH

Ingo Sawilla, TRUMPF Werkzeugmaschinen SE + Co.KG

Dr.-Ing. Mario Schacht, German Institute for Standardization (DIN)

Dr. Peter Schemel, Deloitte GmbH

Kim Marvin Scheurenbrand, Deloitte GmbH

Maximilian Schildt, RWTH Aachen University

David Schirgi, Siemens AG

Dr.-Ing. Miriam Schleipen, EKS InTec GmbH

Dr. Dirk Schlesinger, TÜV AI Lab

Nadine Schlicker, Institut für künstliche Intelligenz in der Medizin, Universitätsklinikum Gießen und Marburg GmbH, Philipps University Marburg

Jun.-Prof. Dr. Thomas Schmid, Martin-Luther University Halle-Wittenberg

Michael-Christian Schmidt, ESKITEC GmbH

Dr. Thomas Schmidt, acatech-Deutsche Akademie der Technikwissenschaften

Christoph Schmidt, German Institute for Standardization (DIN)

Jörg Schmidtke, VIVAVIS AG

Frank Schmiedchen, Vereinigung Deutscher Wissenschaftler e. V. (VDW)

Thorsten Schmitz, EKS InTec GmbH

Jonas Schneider, EFS – Elektronische Fahrwerksysteme GmbH

Detlef Schoepe, Zentrum für Digitalisierung Bundeswehr – Kompetenzzentrum KI

Prof. Dr. Wolfgang M. Schröder, Julius-Maximilians-Universität Würzburg (JMU)

Welf Schröter, Forum Soziale Technikgestaltung

Dr. med. Stephan Schug, Deutsche Gesellschaft für Gesundheitstelematik (DGG) e. V.

Tim Schüßler, Amprion GmbH, Universität Siegen Lehrstuhl Embedded Systems

Jan Fiete Schütte, dimension2 economics & philosophy GmbH

Prof. Dr. Falk Schwendicke, Charité – Universitätsmedizin Berlin

Adrian Seeliger, German Institute for Standardization (DIN)

Prof. Dr. Eberhard K. Seifert, VDW-Studiengruppe Digitalisierung, Senior Fellow IASS-Potsdam

Jan Seitz, Technische Hochschule Wildau (TH Wildau)

Annegrit Seyerlein-Klug, neurocat GmbH

Fatemeh Shahinfar, Institut für angewandte Arbeitswissenschaft e. V. (ifaa)

Prof. Dr. Katharina Simbeck, University of Applied Sciences, Berlin (HTW)

Andreas Skuin, Orban Consulting Holding GmbH

Ariana Sliwa, TÜV AI Lab

Dr. Reiner Spallek, IABG mbH

Dr. Felix Spangenberg, msg systems ag

Philip Sperl, Fraunhofer-Institut für Angewandte und Integrierte Sicherheit (AISEC)

Patrick Spitzer, Deloitte GmbH

Lucas Spreiter, Unetiq GmbH

Prof. Dr. André Steimers, Hochschule Koblenz, Institut für Arbeitsschutz der DGUV

Rosmarie Steininger, CHEMISTREE GmbH

Jannis Steinke, Technische Universität Braunschweig

Mira Stemmer, German Institute for Standardization (DIN)

Dr.-Ing. Patricia Stock, REFA-Institut e. V.

Julia Stoll, Deutsche Akkreditierungsstelle GmbH (DAkkS)

Dr. Christina Strobel, KI Bundesverband e. V.

Dr. Oliver Stuch, Verband der Vereine Creditreform e. V.

Johannes Stürenburg, German Federal Office for Information Security (BSI)

Alexandra Surdina, Deutsche Bahn AG

Ernö Szivek, Deutsche Bundesbank

Dr. Rustam Tagiew, Deutsches Zentrum für Schienenverkehrsforschung (DZSF)

Neal Ternes, ERGO Digital Ventures AG

Sebastian Terstegen, Institut für angewandte Arbeitswissenschaft e. V. (ifaa)

Martin Tettke, Berlin Cert Prüf- und Zertifizierstelle für Medizinprodukte GmbH

Ralph Traphöner, Empolis Information Management GmbH

Holk Traschewski, Your Expert Cluster GmbH

Dr. Volker Treier, Association of German Chambers of Commerce and Industry (DIHK)

Thorsten Trippel, University Tübingen

Kristina Unverricht, German Federal Office for Information Security (BSI)

Dr.-Ing. Thomas Usländer, Fraunhofer Institute of Optronics, System Technologies and Image Exploitation (IOSB)

Dr.-Ing. Mathias Uslar, OFFIS – Institut für Informatik

Tobias van Hasselt, TÜV Nord Mobilität GmbH & Co. KG IFM

Dr.-Ing. Eric MSP Veith, OFFIS – Institut für Informatik

Sharan Vijayagopal, bauforumstahl e. V. (BFS)

Dr. Silvia Vock, Federal Institute for Occupational Safety and Health (BAuA)

Thomas Vollmer, Philips

Dr. Justus von Geibler, Wuppertal Institut für Klima, Umwelt, Energie

Dr. Arndt von Twickel, Federal Office for Information Security (BSI)

Sabine Waechter, Datev eG

Kirsten Wagner, German Technical and Scientific Association for Gas and Water (DVGW)

David Wagner-Stürz, SAMSON AG

Prof. Dr. Siegfried Wahl, Carl Zeiss Vision International GmbH

Prof. Dr. rer. nat. Dr. h.c. mult. Wolfgang Wahlster, Plattform Lernende Systeme / German Research Centre for Artificial Intelligence (DFKI)

Robert Walter, TÜV AI Lab

Dr. Thomas Waschulzik, Siemens Mobility GmbH

Steffen Waurick, German Federal Office for Information Security (BSI)

Dr. Marco Wedel, Technical University of Berlin

Prof. Dr. Dieter Wegener, Siemens AG / German Commission for Electrical, Electronic & Information Technologies of DIN and VDE (DKE)

Eva Weicken, Fraunhofer-Institut für Nachrichtentechnik, Heinrich-Hertz-Institut (HHI)

Felix Wenzel, ERGO Digital Ventures AG

Martin Westhoven, Federal Institute for Occupational Safety and Health (BAuA)

Bernd Wildpanner, Imabicon UG

Dorothea Winter, German Research Centre for Artificial Intelligence (DFKI)

Dr. Johannes Winter, L3S AI Research Center,
formerly: Plattform Lernende Systeme / acatech

Prof. Dr. Mario Winter, German Testing Board e. V. (GTB)

Christoph Winterhalter, German Institute for Standardization (DIN)

Dr. Oliver Wirjadi, Dentsply Sirona

Sebastian Wohlrapp, Field 33 GmbH

Susanna Wolf, Datev eG

Prof. Dr. Stefan Wrobel, Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme (IAIS)

Dr. Alexander Wunderle, infologistix GmbH

Michael Wutz, Vitesco Technologies

Jason YiJunsong, Huawei Technologies

Prof. Dr.-Ing. Sebastian Zaunseder, Dortmund University of Applied Sciences and Arts

Dr. Meike Zehlike, Zalando SE

Dr. Stephan Zidowitz, Fraunhofer-Institut für Digitale Medizin (MEVIS)

Dr. Wolfgang Ziegler, z-rands

Jens Ziehn, Fraunhofer Institute of Optronics, System Technologies and Image Exploitation (IOSB)

Sonja Zillner, Siemens AG

Dr. Bettina Zucker, Drägerwerk AG & Co. KGaA

12 Further members of the working groups

Dr.-Ing. Mohamed Abdelaal, Software AG (SAG)

Katja Anclam, female.vision e. V.

Lisa Auer, RWTH Aachen University

Benedikt Auth, Leistritz Group

Dr. Gergana Baeva, iRights.Lab GmbH

Adam Bahlke, Motor AI

Michael Barth, private person

Jens Bauch, Deutsche Bahn AG

Jessica Bauer, FOM Hochschule für Ökonomie & Management

Stephan Bautz, PricewaterhouseCoopers GmbH

Judit Bayer, University of Münster

Justus Benning, FIR e. V. at RWTH Aachen University

Torsten Berge, Deloitte GmbH

Marc Bergenthal, Brainlab AG

Irvin Bislimi, Aesculap AG

René Böhm, Vitesco Technologies

Andre Bojahr, IAV GmbH

PD Dr. med. Ulrich Bork, Universitätsklinikum Carl Gustav Carus, Technical University Dresden

Dr. Mathis Börner, SAP SE

Thomas Boué, BSA

Robert Brunner, AI4SMB GbR – AI for SMBs in Logistics & Healthcare

Aaron Butler, Universität Luzern

Ralf Casperson, Bundesanstalt für Materialforschung und -prüfung (BAM)

Chih-Hong Cheng, Fraunhofer-Institut für Kognitive Systeme (IKS)

Vasilios Danos, TÜViT

Dr. Werner Daum, Bundesanstalt für Materialforschung und -prüfung (BAM)

Dr. med. Björn Diem, BIOTRONIK SE & Co. KG

Verena Dietrich, imbus AG

Marina Dolokov, FSD Fahrzeugsystemdaten GmbH

Jannis Dörhöfer, TÜV-Verband e. V.

Dr. Patrick Draheim, German Research Center for Artificial Intelligence (DFKI)

Heiko Ehrich, TÜV NORD

Claudia D. Eich, B‘IMPRESS

Kentaro Ellert, PricewaterhouseCoopers GmbH

Jens Elsner, Munich Innovation Labs GmbH

Dr. Rainer Engels, GIZ

Prof. Dr. Kurt Englmeier, Hochschule Schmalkalden

Dr. Nico Erdmann, Deloitte GmbH

Eva Daria Ernst, German Institute for Standardization (DIN)

Dr. Matthias Fabian, private person

Daniel Fehrenbacher, e:fs TechHub GmbH

Jörn Fiedler, German Federal Ministry of Defence (BMVg)

Philip Finkler, Deloitte GmbH

Kerstin Franzl, nexus Institut

Saskia Fruth, Industrie-Anlagen-Betriebs-Gesellschaft

David Fuhr, HiSolutions AG

Christopher Ganz, C. Ganz Innovation Services

Dr. Jens Gayko, German Commission for Electrical, Electronic & Information Technologies of DIN and VDE (DKE)

Prof. Dr. Raimund Geene, Alice Salomon Hochschule Berlin

Regina Geierhofer, Siemens Healthcare GmbH

Sebastian Giera, Robert Bosch GmbH

Dr. Patrick Gilroy, TÜV-Verband e. V.

Lea Gimpel, GIZ

Dr. Robert Ginthör, Know-Center GmbH

Rebekka Görge, Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme (IAIS)

Dr. Alexander Goschew, German Institute for Standardization (DIN)

Dr. Maximilian Grabowski, BASt

Marian Gransow, VIVE-MedTech GmbH

Stephan Griebel, Siemens Mobility GmbH

Claudia Großmann, Modis GmbH

Tanja Hagemann, Deutsche Telekom AG

Andreas Halbleib, B. Braun Gruppe

Prof. Anselm Haselhoff, Hochschule Ruhr West

Dr. Vahid Hashemi, Audi AG

Thomas Heckel, Bundesanstalt für Materialforschung und -prüfung (BAM)

Manfred Hefft, Domino Deutschland GmbH

Dr. Jens Heidrich, Fraunhofer-Institut für Experimentelles Software Engineering (IESE)

Mathias Heiles, LIME medical GmbH

Martina Heim, Alcon

Jana Heinrich, Fraunhofer-Institut für Experimentelles Software Engineering (IESE)

Paul Hellwig, BG Kliniken IT-Services GmbH

Andreas Hepfner, Neo Q Quality in Imaging GmbH

Karol Tatiana Puscus Hernandez, RWTH Aachen University

Dr.-Ing. Stefan Hillmann, Technische Universität Berlin

Karl Peter Hoffmann, Stadtwerke Sindelfingen

Reiner Hofmann, Universität Bayreuth, Medizincampus Oberfranken

Christoph Hohenberger, retorio GmbH

Dr. Johannes Hüdepohl, Berufsgenossenschaft Energie Textil Elektro Medienerzeugnisse – BG ETEM

Marc Jopek, Lyniate

Juliane Jungk, Freudenberg & Co. KG

Dr. Michael Karner, SETLabs Research GmbH

Sven Kasan, Digithurst Bildverarbeitungssysteme

Klaus Kaufmann, Mittelstand 4.0 Kompetenzzentrum eStandards

Dr. Hubert B. Keller, ci-tec GmbH

Dr. Christian Kellermann, Federation of German Scientists (VDW)

Michael Kieviet, innotec GmbH

Dorian Knoblauch, Fraunhofer-Institut für Offene Kommunikationssysteme (FOKUS)

Dr. Gesine Knobloch, Bayer AG

Sabine Knör, Atos

Johannes Koch, German Commission for Electrical, Electronic & Information Technologies of DIN and VDE (DKE)

Philipp Koch, German Research Centre for Artificial Intelligence (DFKI)

Michael Kolain, Deutsches Forschungsinstitut für öffentliche Verwaltung (FÖV)

Dr. Sergii Kolomiichuk, Fraunhofer-Institut für Fabrikbetrieb und -automatisierung (IFF)

Roman Konertz, FernUniversität in Hagen

Dr.-Ing. Dietmar Köring, Arphenotype

Stephan Krähnert, German Association of the Automotive Industry (VDA)

Dr. rer. nat. Joachim Krois, Charité – Universitätsmedizin Berlin

Christian Kruschel, IAV GmbH

Prof. Dr. Kai-Uwe Kühnberger, University of Osnabrück

Mark Küller, TÜV-Verband e. V.

Dr. Kai Kümmel, private person

Philipp Lämmel, Fraunhofer-Institut für Offene Kommunikationssysteme (FOKUS)

Sebastian Land, Old World Computing GmbH

Claus Lang, Kodex AI GmbH

Yves Leboucher, Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ), formerly: German Commission for Electrical, Electronic & Information Technologies of DIN and VDE (DKE)

Dr. Andreas Lemke, mediaire GmbH

Johann Letnev, JUMO GmbH & Co. KG

Ulli Leucht, PricewaterhouseCoopers GmbH

David Lewenko, Deloitte GmbH

Matthias Lieske, Hitachi Europe GmbH

Thomas Linner, OTH Regensburg

Alina Lorenz, IT-Systemhaus der Bundesagentur für Arbeit

Mihai Maftei, German Research Centre for Artificial Intelligence (DFKI)

Hans-Christian Mangelsdorf, Federal Foreign Office

Angelina Marko, Bitkom e. V.

Dr.-Ing. Erik Marquardt, VDI Verein Deutscher Ingenieure e. V.

Björn Matthias, ABB AG

PD Dr. Matthias May, Universitätsklinikum Erlangen

Benjamin Meier, Curalie GmbH

Andreas Meisenheimer, Bundeswehr

Christian Meyer, msg-systems AG

Stephan Mietke, Bundesverband deutscher Banken

Olaf Minkwitz, Marelli Automotive Lighting

Dr. Klaus Möller, DEFINO Institut für Finanznorm AG

Dr. Julia Maria Mönig, Bonn University

Bernhard Mühlbauer, Energie Baden-Württemberg AG (EnBW)

Dr. Frank Müller, Heidelberg Engineering GmbH

Wolfgang Müller, Zentralverband der Augenoptiker und Optometristen

Corinna Mutter, SPECTARIS – Deutscher Industrieverband für Optik, Photonik, Analysen- und Medizintechnik e. V.

Florian Neumeier, M3i Industrie-in-Klinik-Plattform

Dr. Jens Niederhausen, National Metrology Institute (PTB)

Franziska Noack, private person

Jan Noelle, RKiSH gGmbH

Alexander Nollau, German Commission for Electrical, Electronic & Information Technologies of DIN and VDE (DKE)

Prof. Dr. Dirk Nowotka, Christian-Albrechts-Universität zu Kiel

Karolina Ochs, Christian-Albrechts-Universität zu Kiel

Stefan Otterbach, German Federal Ministry of Defence (BMVg)

Dr. Daniel Paulus, Acosu

Juliane Pfeil, Technische Hochschule Wildau (TH Wildau)

Dr. Christian Piovano, ZF Friedrichshafen AG

Dr. Axel Plinge, Fraunhofer-Institut für Integrierte Schaltung (IIS)

Bernd Püttmann, TÜV NORD CERT GmbH

Dr. Frederic Raber, Federal Office for Information Security (BSI)

Myriam Raboldt, TU Berlin

Dr. Hans Rabus, National Metrology Institute (PTB)

Felix Rau, Köln University

Lukas Rauh, Fraunhofer-Institut für Produktionstechnik und Automatisierung (IPA)

Martin Reich, MORE THAN CAPITAL

Dr. Alexander Reiprich, KARL STORZ Endoskopie Berlin GmbH

Ina Reis, Senatskanzlei Hamburg, Amt für IT und Digitalisierung

Luca Rettenberger, Karlsruhe Institute for Technology (KIT)

Christian Richter, Verwaltungs-Berufsgenossenschaft (VBG)

Dr. Patrick Riordan, Siemens AG

Renato Rodrigues, DB Netz

Maximilian Rohleder, Friedrich-Alexander-Universität Erlangen-Nürnberg

Christian Rudolf, MHP

Ingo Rütten, Strategieberatung Zielwerk GmbH

Peter Salathe, m.Doc GmbH

Sophia Saller, SMF

Friedrich Sanzi, Leuze electronic GmbH + Co. KG

Christian Schaaf, Universitätsklinikum Heidelberg

Daniel Schäfer, Hermann Bock GmbH

Stefan Schaffer, German Research Centre for Artificial Intelligence (DFKI)

Michaela Schierholz, German Institute for Standardization (DIN)

Dr. Jasmine Schirmer, Carl Zeiss Meditec AG

Hans-Dieter Schmees, Verein Deutscher Werkzeugmaschinenfabriken e. V. (VDW)

Andreas Schmidt, ZF Friedrichshafen AG

Jonas Schmidt, ZF Friedrichshafen AG

Christian Schmitz, Novar GmbH a Honeywell Company

Dr. med. ETH Rüdiger Schmitz, Universitätsklinikum Hamburg-Eppendorf

Dr.-Ing. Fabian Schnabel, Fachverband des Tischlerhandwerks Nordrhein-Westfalen

Frank Schneider, TÜV-Verband e. V.

Mark Schutera, ZF Friedrichshafen AG

Daniel Schwabe, National Metrology Institute (PTB)

Dr. Joachim Seeler, HSP Hamburg Invest GmbH

Roman Senderek, FIR e. V. at RWTH Aachen University

Aydin Enes Seydanlioglu, Robert Bosch GmbH

Dr. Georgy Shakirin, Carl Zeiss Meditec AG

Ankur Sharma, Bayer AG

Kris Shrishak, Irish Council for civil liberties

Tomasz Soltysinski, QuIP GmbH

Georg Peter Sotiriadis, Phantasma Labs GmbH

Dirk Spaltmann, Bundesanstalt für Materialforschung und -prüfung (BAM)

Florian Stark, Industrial Analytics IA GmbH

Christina Stathatou, Kugler Maag CIE

Jan Stodt, Hochschule Furtwangen (HFU)

Christian Stohs, Union Investment

Prof. Dr.-Ing. habil. Sascha Stowasser, Institut für angewandte Arbeitswissenschaft e. V. (ifaa)

Volker Sudmann, mdc medical device certification GmbH

Dima Taleb, TÜV Rheinland

Dr.-Ing. Nikolay Tcholtchev, Fraunhofer-Institut für Offene Kommunikationssysteme (FOKUS)

Dr. habil. Florian Thiel, National Metrology Institute (PTB)

Heike Thomas, UL International Germany GmbH

Jack Thoms, German Research Centre for Artificial Intelligence (DFKI)

Verena Till, Think Tank iRights.Lab

Hauke Timmermann, eco Verband der Internetwirtschaft e. V.

Mario Tokarz, RightMinded AI GmbH

Kevin Trelenberg, Hochschule Ruhr West

Merle Uhl, Bitkom e. V.

Dr. Thomas Unger, KraussMaffei Extrusion GmbH

René Urban, Unitransferklinik Lübeck

Bhaskar Vanamali, Kugler Maag CIE

Sonja Verschitz, Digital Humanities – Konzept und Strategie: Daten – Information – Wissen

Annette von Wedel, female.vision e. V.

Ronny Wegner, PAUL HARTMANN AG

Christoph Wehner, Otto-Friedrich-Universität Bamberg

Prof. Dr. Joh Wilh Weidringer, private person

Reinhard Weissinger, ISO

Frank Werner, Software AG (SAG)

Lucas Weyrich, Robo Test

Dr. Sebastian Wieczorek, SAP SE

Rick Wilming, TU Berlin

Fabian Witt, MATHEMA GmbH

Sebastian Witte, Bundesverband Digitale Wirtschaft

Dr. Nicole Wittenbrink, VDI/VDE Innovation + Technik GmbH

Georg Woditsch, Alexianer GmbH

Thorsten Wujek, SALT AND PEPPER Technology

Semih Yalcin, TakeAway Express GmbH (Lieferando)

Marc Zeller, Siemens AG

Jing Zhang, Huawei Technologies

Klaus Ziegler, International Association of Conference Interpreters (AIIC)

13
Annex

13.1 Annex Artificial Intelligence Act (AI Act)

The following Table 16 provides a brief description of the contents of and relevance to the AI Act for the EU laws presented in Figure 6. In addition, further details such as the type of legislation, related laws at German level and also the status of legislation are listed.

Table 16: EU laws with particular relevance to the AI Act

1a:

EU Charter of Fundamental Rights

Full title: Charter of Fundamental Rights of the European Union

Status: legally binding since 1 December 2009

Description:

The Charter of Fundamental Rights of the European Union codifies fundamental and human rights. In six Chapters (Dignity, Freedoms, Equality, Solidarity, Citizens' Rights and Justice) the Charter summarizes general human and civil rights and economic and social rights in one document. The Charter contains some essential principles to which the European legislator in particular must adhere. In 50 Articles, comprehensive rights are recognized, for the enforcement of which not only the European Court of Justice in Luxembourg, but also all national judges – as Union judges, so to speak – are responsible. Article 1 of the Charter, like Article 1(1) of the Basic Law of the Federal Republic of Germany, states: "Human dignity is inviolable." It also regulates areas of protection that are not explicitly mentioned in the German Basic Law, such as the protection of personal data, the right to education, the rights of children, people with disabilities and the elderly, the right to good administration or the guarantees in labour law. Furthermore, consumer protection, inviolability of the home, telecommunications secrecy, "dignified working conditions" and free employment services are guaranteed. In addition, the Charter is steeped in anti-discrimination. Art. 21 states "Any discrimination based on any ground such as sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation shall be prohibited."

The Charter of Fundamental Rights also applies to AI applications and is the basis for technical realization to avoid unintentional discrimination.

1b:

Product liability directive

Full title: Council Directive 85/374/EEC of 25 July 1985 on the approximation of the laws, regulations and administrative provisions of the Member States concerning liability for defective products

Status: entered into force, implemented in Germany as the Product Liability Act (Produkthaftungsgesetz – Gesetz über die Haftung für fehlerhafte Produkte).

Description:

The Commission is concerned that the opacity and complexity, as well as the high degree of autonomy, of some AI systems may make it difficult for injured parties to prove a product's defectiveness or fault, as well as the causal link to the harm. It could also be uncertain whether and to what extent national rules on strict liability (e.g., for dangerous activities) apply to the use of AI-supported products or services.

The Commission intends to counter these risks, if necessary, through various measures such as easing the burden of proof or strict liability of the manufacturer.

1c:

Occupational safety and health framework directive

Full title: Council Directive 89/391/EEC of 12 June 1989 on the introduction of measures to encourage improvements in the safety and health of workers at work

Status: entered into force, implemented in Germany as the Occupational Health and Safety Act (Arbeitsschutzgesetz).

Description:

The aim of the Directive is to create a standardized scheme for all employees with regard to health and safety. Under this law, employers are required to take appropriate preventive measures to improve safety and health. One of the Directive's key points is risk assessment, which highlights the following topics, among others:
Identification of hazards in the workplace and their harmful effects

Appropriate measures to combat potential risks

Procedures for documentation

1d:

Machinery Directive (or Machinery Regulation)

up to now an EU Directive

Full title: Directive 2006/42/EC of the European Parliament and of the Council of 17 May 2006 on machinery and amending Directive 95/16/EC

Status: entered into force, implemented in Germany as the Product Safety Act (Produktsicherheitsgesetz)

in the future it will be an EU Regulation:

Full title: Regulation of the European Parliament and of the Council on machinery products

Status: planned

Description:

The EU Machinery Directive regulates uniform requirements for machinery and parts of machinery for a uniform level of protection to prevent accidents when placing them on the market. In Germany, the Directive has been incorporated into the Product Safety Act (ProdSG) and the associated Machinery Ordinance (9. ProdSV). The following requirements must be implemented (selection):
The machine must be designed to be mechanically and electrically safe and functional safety (e.g. safe control circuits) must be implemented,

at the time of placing on the market, the machine is safe and safe operation is ensured,

safety devices or protective devices of the machine cannot be easily by-passed,

conformity assessment procedures with risk assessment (Section 158 ff.) are carried out,

after successful assessment, the declaration of conformity and affixing of the CE mark takes place,

technical documentation and operating instructions that clearly draw the attention of users and operators of the machine to the identified residual risks are prepared.

Where AI components are installed in or for "machinery", the requirements of the Machinery Directive apply. However, specific considerations on risks from AI and related measures are not included there. As with the sectoral harmonization regulations (e.g. Medical Device Regulation), the CE mark has been awarded via the Machinery Directive up to now.

On April 21, 2021, the European Commission presented a proposal to transform the Machinery Directive into a Machinery Regulation (Proposal for a Regulation of the European Parliament and of the Council on Machinery Products, Brussels, April 21, 2021, [346] 202 final 2021/0105 (COD)), which is embedded in the New Legislative Framework (NLF, 768/2008/EC). It seeks full compatibility with the AI Act, explicitly takes up the term "artificial intelligence," and identifies high-risk systems comparable to the AI Act, including "machinery embedding AI systems ensuring safety functions."

1e:

Medical Device Regulation (MDR) as an example of sector-specific regulations for the safety of products in the respective areas of application

Full title: Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC

Status: entered into force in 2017

Description:

The MDR regulates the approval and operation of medical devices. It sets key requirements for their safety and efficacy, including requirements for the development process of medical devices and all subsequent measures to ensure safe manufacture, commissioning and operation. The MDR does not contain specific requirements for AI-based systems that constitute or are a component of a medical device. The AI Act seeks to address this gap in sectoral harmonization legislation by establishing basic requirements for AI systems in a horizontal approach. Consistency between the AI Act (horizontal) and the MDR (sectoral) should be ensured to enable and not hinder the implementation of AI-based medical devices.

2a:

General Data Protection Regulation (GDPR)

Full title: Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

Status: entered into force in 2016

Additional German laws: Bundesdatenschutzgesetz (Federal Data Protection Act), Landesdatenschutzgesetze (data protection acts of the Länder)

Description:

The GDPR regulates the processing of personal data. Art. 5, 24, 25 and 32 contain responsibilities, the preparation of a data protection impact assessment (risk assessment) and requirements for data protection-friendly and secure technology and organization (including pseudonymization and encryption).

For automated decision-making, e.g., from machine learning (ML) models involving individuals, the following passage is critical: “Where personal data relating to a data subject are collected from the data subject, the controller shall ... provide the data subject with all of the following information: ...the existence of automated decision-making ... and ... meaningful information about the logic involved”.

To determine the risks to data subjects, data protection supervisory authorities across Europe have agreed on nine criteria:

1. assessment or classification,
2. automatic decision-making,
3. systematic supervision,
4. confidential or highly personal data,
5. large-scale data processing,
6. synchronization or merging of datasets,
7. data on vulnerable data subjects,
8. innovative use or application of new technological or organizational solutions,
9. data subjects are prevented from exercising a right or using a service or performing a contract.

The aforementioned risk criteria and their assessment are relevant when using an AI where personal data is used. Meaningful information about the logic used must be available, i.e. transparency about the origin of the decision of an AI. In the “Hambach Declaration on Artificial Intelligence” [43] the German data protection supervisory authorities make a concrete statement on the requirements of the GDPR with regard to AI.

2b:

Network Information Security (NIS) Directive

Full title: Directive (EU) 2016/1148 of the European Parliament and of the Council of 6 July 2016 concerning measures for a high common level of security of network and information systems across the Union

Status: entered into force in 2016

Description:

The aim of the Directive is to achieve a uniformly high level of security of network and information systems across the EU through increased cybersecurity capacity at national level, enhanced cooperation at EU level and obligations for operators of essential services and digital service providers, minimum security requirements for risk provisioning and maintenance of essential services, and reporting requirements. Sectors have been defined as critical infrastructure, such as energy, transport, health and digital infrastructure, as well as sanctions. The NIS Directive has been implemented in Germany with the IT Security Act 1 and 2 (IT Sicherheitsgesetz 1 und 2).

2c:

Cybersecurity Act (CSA)

Full title: Regulation (EU) 2019/881 of the European Parliament and of the Council of 17 April 2019 on ENISA (the European Union Agency for Cybersecurity) and on information and communications technology cybersecurity certification and repealing Regulation (EU) No 526/2013 (Cybersecurity Act)

Status: entered into force in 2019

Description:

The aim of the Cybersecurity Act is to establish uniform regulations throughout the EU to strengthen IT security for information and communications technology (ICT) systems, services and processes.

Core elements of the CSA are a permanent mandate for the European Union Agency for Cybersecurity (ENISA) and the introduction of a uniform European certification framework for ICT products, services and processes. These are to be certified according to defined security levels as “low”, “medium” and “high” according to various provisions.

Potential IT security risks from AI are not specifically described or considered. It is recommended to check to what extent additions might be necessary. AI-based products, however, are subject to IT system cybersecurity requirements and must implement them accordingly.

2d:

Cyber Resilience Act (CRA)

Status: planned

Description:

The CRA governs cybersecurity requirements for a wide range of digital products and associated ancillary services. The subjects of the Act are tangible digital products and non-embedded software over their entire life cycle. Thus, the Act covers hardware and software equally.

The planned Act defines the following three main objectives:
Ensure a consistently high level of cybersecurity for digital products and ancillary services

Increase the transparency of cybersecurity features

Create a level playing field for providers of digital products and ancillary services

3a:

Data Governance Act (DGA)

Full title: Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European data governance and amending Regulation (EU) 2018/1724 (Data Governance Act)

Status: entered into force in 2022

Description:

The DGA is intended to create impetus for better use of valuable data throughout Europe. It is just as much about public sector data as it is about data of companies.

Examples include environmental data from smart-home devices that could help combat climate change, as well as a greater use of health data for research purposes.

The regulation is intended to facilitate access to both personal data and non-personal data. It supplements the EU's Open Data Directive adopted last year. The regulation is intended to help make data available in a simple and legally secure manner.

The regulation also creates a legal framework for "data intermediaries". These are neutral intermediaries designed to facilitate exchanges between data sources and interested parties. Individuals whose personal data is used, on the other hand, are to be able to organize themselves into data cooperatives in the future. The EU Commission also wants to make it easier to donate data for charitable purposes under the heading of data altruism, as is already the case with the German Corona data donation app.

The DGA does not seek to grant, modify, or eliminate substantial rights to the access and use of data.

3b:

Digital Services Act (DSA)

Full title: Proposal for a Regulation of the European Parliament and of the Council on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC

Stand: Planned, shortly before completion of the legislative process; validity expected from 1 January 2024 at the latest.

Description:

The DSA is intended as a kind of "charter of the Internet" to protect the digital space against the dissemination of illegal content and to guarantee the fundamental rights of users. It aims to prevent the spread of hate speech and disinformation, strengthen consumer protection online, and create transparency about how digital services work.

The DSA follows the principle that what is illegal offline must also be illegal online.

Essentially, online platforms, including social media and marketplaces, must take measures to protect users from illegal content, goods and services. The DSA will apply to all online intermediaries offering services in the EU, but very large online platforms ("VLOPs") and very large online-search engines ("VLOSEs"), i.e. services with more than 45 million active users in the EU, will be subject to stricter requirements than micro and small businesses, which are exempt from some of the obligations.

Upon request by the competent authority, particularly large online platforms must provide the competent authority with access to data necessary to monitor compliance with the DSA.

3c:

Digital Markets Act

Full title: Regulation of the European Parliament and of the Council on contestable and fair markets in the digital sector (Digital Markets Act)

Status: planned

Description:

A few very large online platforms account for a very large share of the digital economy in the EU. Their economic power and control over entire platform ecosystems often make it impossible for competitors or new entrants to compete. The Digital Markets Act identifies opportunities for regulating large, "gatekeeper" online platforms.

Pursuant to Art. 19 of the DMA, the European Commission may also request access to databases and algorithms of companies by simple request for information or by way of a decision, and request explanations in this regard.

3d:

Data Act

Full title: Regulation of the European Parliament and of the Council on harmonised rules on fair access to and use of data (Data Act)

Status: planned

Description:

Through "networked" products and services in the Internet of Things (IoT), data is generated on a significant scale and is of considerable value, for example when driving one's own car or controlling one's own heating. The vast majority of this data is currently either not used at all or only a few very large companies benefit from it. The Data Act aims to ensure a fairer distribution of the value created by data and to promote competitiveness and innovation in the European Union through the reusability of data. The new rules are expected to make more data available for reuse and are expected to generate an additional € 270 billion in GDP by 2028.

To this end, the draft regulation creates a new right to access and use of data for certain private and public actors.

3e:

European Health Data Space (EHDS) as an example of sector-specific regulations regarding access to data in the respective areas of application

Full title: Regulation of the European Parliament and of the Council on the European Health Data Space

Status: planned

Description:

The EHDS is intended to regulate access to health data. On the one hand, this concerns measures for individuals to control their own data. On the other hand, it promotes the use of health data to enable better medical care, especially for research, innovation, and policy-making. In doing so, it seeks to fully exploit the potential for exchange, use and reuse of data under the standard of interoperability, but also while ensuring secure access.

With regard to the planned AI Act, it represents a key element for AI or machine learning-based approaches to be implemented in healthcare. Both regulations contain requirements with regard to accessing or handling the data required for the AI systems, which in the medical environment initially often represent personal data and must therefore be pre-processed in a suitable manner, in particular using methods of anonymization or pseudonymization.

Example of medical devices

AI-based medical devices are among the areas that will have to comply with two harmonization requirements after the planned AI Act becomes effective. In addition to the planned AI Act, another EU regulation is the Medical Device Regulation (MDR) 2017/745. Medical devices are categorically considered to be high-risk devices within the meaning of the AI Act pursuant to Art. 6(1) and Annex II as soon as they are subjected to a conformity assessment in accordance with the MDR. The MDR itself contains its own risk classification, which includes classes I, IIa, IIb and III, with severity factored in for potential harm to patients or users. According to the application of Classification rule 11 (see Annex VIII of the MDR), virtually all stand-alone medical software is classified at least in Class IIa, and in the case of higher hazard potential also IIb or III. Due to the requirements of the MDR, the product subsequently undergoes a supervised conformity assessment procedure in accordance with the MDR and is thus a high-risk system within the meaning of Art. 6 para. 1 of the draft AI Act.

This means that the relevant requirements of the AI Act apply, e.g. with regard to information security (cybersecurity), implementation of a risk management system, post-market surveillance, reporting system, technical documentation, labeling, QM system and entry in a product database. The MDR also requires the same. However, the two regulations differ from each other in some points and contain inconsistencies that should be eliminated so that products can be placed on the market in accordance with both regulations. For example, the quality management system according to DIN EN ISO 13485:2021 [381] to be applied for compliance with the essential safety and performance requirements of the MDR is in principle compatible with the requirements of the draft of the planned AI Act. However, the following requirements from this draft Act are not considered:

→ Specific procedure for managing the data required to train the device before and for the purpose of placing it on the market
→ Adaptation of the procedure for communication with market authorities: Access to data
→ Adaptation of the design and development process to meet Annex VI requirements (e.g., AI system training, human oversight)
→ Adaptation of the risk management system. According to Art. 9(8), one of the factors to be considered is whether the high-risk AI system is likely to be accessible to or have an impact on children.

In addition, there is a separate product database (Eudamed) in the MDR according to Art. 33, which is used for a variety of purposes. This includes the registration as well as a basic description of products including information about the manufacturer and other relevant economic operators, performance records associated with the product (including clinical trials) as well as collected information regarding vigilance and market surveillance. It remains unclear whether the product database required under Art. 60 of the AI Act is already given by the Eudamed database or whether it is to be a stand-alone database. In the latter case, this would mean a duplication of effort regarding the maintenance of the data with the additional risk of inconsistencies in the reporting of incidents, e.g. due to different requirements.

Another challenge is conflicting risk management requirements. While the MDR allows a risk-benefit balance (cf. Art. 2 No. 24 MDR), according to which a medical device may be placed on the market if the benefits of a product outweigh the associated risks (potential harmfulness, see Annex I No. 8 MDR), the draft of the planned AI Act follows an ALAP (as low as possible) approach, according to which risks must be mitigated as far as possible, regardless of the benefits. Since conformity with both harmonization rules must be ensured, this would mean that the stricter rules would always have to be observed. However, as in the case of risk management, the requirements of the proposed AI Act may not be appropriate for the specific application. In the field of medical devices, the balancing of risks and benefits is a central feature for conformity assessment.

Further overlaps and inconsistencies in requirements can lead to problems in the approval process. According to Recital 63 and Art. 43(3) of the draft of the planned AI Act, it should be sufficient – in order to avoid duplication – for high-risk AI systems to undergo only one conformity assessment procedure under an applicable provision listed in Annex II. However, this assumes that the notified body is also certified for the AI Act. However, if that notified body does not seek certification to the AI Act, another notified body must be used to monitor compliance with the AI Act. Since the number of notified bodies certified to the MDR is still very low, an additional restriction is added here. Combined with the skills shortage in this field, which will gain further complexity with the incorporation of AI, there are likely to be significant bottlenecks.

From the perspective of medical technology, therefore, the AI Act is more likely to act as a brake on innovation than to promote it. Medical device manufacturers may only use AI to improve the safety, performance and efficacy of devices (Annex I, Chapter I MDR). The results of a company survey published by SPECTARIS (see "First assessment of German manufacturers of medical devices after the EU Medical Device Regulation (MDR) came into force"; ed. by: Deutscher Industrie- und Handelskammertag e. V., MedicalMountains GmbH, SPECTARIS. Deutscher Industrieverband für Optik, Photonik, Analysen- und Medizintechnik e. V.; Berlin, Tuttlingen; April 2022) show a significant extension of conformity assessment procedures involving a notified body of 45 % on average. In risk class III, the duration of conformity assessment procedures has even more than doubled (101 %). In addition, numerous products are already being withdrawn from the market, many innovation products are on hold, and most existing products have not yet been transferred to the MDR.

Not many manufacturers will take the risk of further delay due to additional requirements from an AI regulation.

Another inconsistency arises from the application of the planned AI Act in parallel with the requirements of the GDPR. Article 64(1) of the draft AI Act requires "access to data and documentation in the context of their activities, the market surveillance authorities shall be granted full access to the training, validation and testing datasets used by the provider, including through application programming interfaces ('API') or other appropriate technical means and tools enabling remote access" for market surveillance authorities. It cannot be ruled out that confidential patient data will be used to train, validate, and test an AI system. Making them accessible remotely is contrary to the rules of the GDPR, which defines health data as personal data requiring special protection. So either one regulation or the other is violated. This would mean a further regulation on the handling of data in addition to the planned EU Data Act.

13.2 Annex Language technologies

Existing special standards and specifications in the field of language technology in the broad sense at national, European and international level

Design

→ ISO 9241 Ergonomics of human-system-interaction – Part 110: Dialogue principles
  - Last reviewed and confirmed in 2018. → Now under review
  - DIN EN ISO 9241-110:2020 → Published in May 2020

→ ISO 9241 Ergonomics of human-system-interaction – Part 154: Interactive voice response (IVR) applications)
  - Last reviewed and confirmed in 2020. → Now confirmed

→ ISO 9241 Ergonomics of human-system interaction – Part 11: Usability: Definitions and concepts
  - Published on 2018-04-04

→ ISO 9241-210 Ergonomics of human-system interaction – Part 210: Human-centred design for interactive systems
  - Published on 2019-07-04

→ ISO 9241 Ergonomics of human-system interaction – Part 171: Guidance on software accessibility
  - International standard confirmed on 2018-12-08

→ AS 5061
  - Withdrawn 2019

Voice interaction

→ ETSI ES 202 076 V2.1.1

→ ISO/IEC 30122 Information technology – User interfaces – Voice commands – Part 1: Framework and general guidance
  - ISO/IEC 30122-1:2016 → 08-2016

→ ISO/IEC 30122 Information technology – User interfaces – Voice commands – Part 2: Constructing and testing
  - ISO/IEC 30122-2:2017 → 02-2017
  - 15.01.2022 Under systematic review

→ ISO/IEC 30122 Information technology – User interfaces – Voice commands – Part 3: Translation and localization
  - ISO/IEC 30122-3:2017 → 02-2017
  - 15.01.2022 Under systematic review

→ Voice Control API (VOCAPI)

→ Web Speech API
  - Draft Community Group Report, 18 August 2020

NLP

→ ISO 24617-2 Language resource management – Semantic annotation framework (SemAF) – Part 2: Dialogue acts
  - Under review, it will be replaced by ISO/DIS 24617-2
  - ISO 24617-2:2020 Language resource management – Semantic annotation framework (SemAF) – Part 2: Dialogue acts → 02.12.2020
  - Speech Recognition Grammar Specification (SRGS); '16 March 2004
  - Semantic Interpretation for Speech Recognition (SISR); 5 April 2007

→ **ISO standards:**
Foundational and terminological standards:
  - ISO/IEC 2382: Information technology – Vocabulary
  - ISO/IEC 22989:2022: Information technology – Artificial intelligence – Artificial intelligence concepts and terminology
  - ISO/IEC 24029-2 Information technology – Artificial Intelligence (AI) – Assessment of the robustness of neural networks – Part 2: Methodology for the use of formal methods

Natural language data

→ ISO 5127: Information and documentation – Foundation and vocabulary

ISO/TC 37 projects:

→ ISO 639 series: Codes for the representation of names of languages
→ ISO/TR 20694: A typology of language registers
→ ISO/TR 21636: Identification and description of language varieties
→ ISO 24611: Language resource management – Morpho-syntactic annotation framework (MAF)
→ ISO 24612: Language resource management – Linguistic annotation framework (LAF)
→ ISO 24614 series: Language resource management – Word segmentation of written texts
→ ISO 24615 series: Language resource management – Syntactic annotation framework (SynAF)
→ ISO 24617 series (especially parts 2 and 4): Language resource management – Semantic annotation framework (SemAF)
→ ISO 24624: Language resource management – Transcription of spoken language
→ ISO 24619: Language resource management – Persistent identification and sustainable access (PISA)
→ ISO 20539: Translation, interpreting and related technology – Vocabulary
→ ISO 17100: Translation services – Requirements for translation services

ISO/TC 159 PROJECTS:

→ ISO 24551: Ergonomics – Accessible design – Spoken instructions of consumer products
→ ISO 9241-154: Ergonomics of human-system interaction – Part 154: Interactive voice response (IVR) applications
→ ISO/TR 19358: Ergonomics – Construction and application of tests for speech technology

ITU-T standards:

Projects from SG16 “Multimedia coding, systems and applications”:

→ ITU-T F.745: Functional requirements for network-based speech-to-speech translation services
→ ITU-T F.746.5: Framework for a language learning system based on speech and natural language processing (NLP) technology
→ ITU-T F.746.10: Architecture for a spontaneous dialogue processing system for language learning
→ ITU-T H.625: Architecture for network-based speech-to-speech translation services
→ ITU-T H.862.5 (ex F.EMO-NN): Emotion enabled multi-modal user interface based on artificial neural network
→ ITU-T F.746.11 (ex F.IQAS-INT): Interfaces for intelligent question answering system
→ ITU-T F.AI-FASD: Framework for audio structuralizing based on deep neural network (see work programme)
→ ITU-T F.AI-SCS: Use cases and requirements for speech interaction of intelligent customer service (see work programme)
→ ITU-T F.REAIOCR: Requirements and evaluation methods for AI-based optical character recognition service (see work programme)
→ ITU-T F.AI-RMCDP: Requirements of multimedia composite data preprocessing (see work programme)
→ ITU-T FSTP-ACC-AI: Guideline on the use of AI for ICT accessibility (see work programme)

Quality assessment

Projects from SG12 “Performance, quality of service and quality of experience”:

→ ITU-T P.1130: Subsystem requirements for automotive speech services
→ ITU-T P.1140: Speech communication requirements for emergency calls originating from vehicles
→ ITU-T P.1150: In-car communication audio specification

→ ITU-T P.59: Artificial conversational speech
→ ITU-T P.85: A method for subjective performance assessment of the quality of speech voice output devices
→ ITU-T P.807: Subjective test methodology for assessing speech intelligibility
→ ITU-T Rec. P.851: Subjective quality evaluation of telephone services based on spoken dialogue systems
→ ITU-T P.Sup24: Parameters describing the interaction with spoken dialogue systems

W3C Community Groups

→ Voice Interaction Community Group
- JSON Representation of Semantic Information → last modified: February 12, 2019
- Intelligent Personal Assistant Architecture → Architecture and Potential for Standardization Version 1.0 → Last modified: March 24, 2020
- Intelligent Personal Assistant Architecture → Architecture and Potential for Standardization Version 1.2 → Last modified: July 19, 2021

→ Conversational Interfaces Community Group
- Dialogue Manager Programming Language (DMPL) → Final Community Group Report 13 April 2020
- DM Script (DMS) → Final Community Group Report 13 April 2020

→ Voice Assistant Standardisation Community Group → nothing new
→ The Voice Browser Working Group
- Closed on 2015-10-12.

→ Multimodal Interaction Working Group
- Closed in February 2017

→ https://www.w3.org/community/mqmcg/
→ https://www.astm.org/workitem-wk46396

W3C standards

→ Voice Extensible Markup Language (Voice XML) Version 2.0
- W3C Recommendation 16 March 2004
- VoiceXML Version 3.0 → W3C Working Draft 16 December 2010

→ Speech Synthesis Markup Language (SSML) Version 1.1
- W3C Recommendation 7 September 2010

→ Pronunciation Lexicon Specification (PLS) Version 1.0
- W3C Recommendation 14 October 2008

→ EMMA: Extensible MultiModal Annotation markup language
- W3C Recommendation 10 February 2009
- EMMA: Extensible MultiModal Annotation markup language Version 2.0 → W3C Working Group Note 2 February 2017

Other projects

→ COMPRISE: D5.1 Data protection and GDPR requirements

Other regulations

→ Interstate Media Treaty (Medienstaatsvertrag – MStV) → new regulations on firms or technologies that serve as intermediaries to online media services. → 7 November 2020 (https://www.die-medienanstalten.de/fileadmin/user_upload/Rechtsgrundlagen/Gesetze_Staatsvertraege/Interstate_Media_Treaty_en.pdf
→ European Data Protection Board: „Guidelines 02/2021 on virtual voice assistants“ → February 2021 (https://edpb.europa.eu/our-work-tools/our-documents/guidelines/guidelines-022021-virtual-voice-assistants_en)

Associations

→ Open Voice Network (OVON)

De facto standards:

[tokenization, PoS tagging, dependency parsing] Universal Dependencies guidelines
[language identification] Formats and metrics of the NIST LRE challenge series (documentation)
[speaker detection] Formats and metrics of the NIST SRE challenge series (documentation)
[machine translation] .sgm format (based on SGML), OPUS data formats (documentation)
[machine translation] NIST and BLEU evaluation metrics, sacreBLEU evaluation tool
[automatic summarization] ROUGE evaluation metrics
[word embeddings] word2vec format (space-separated), GloVe format (without header), fastText binary format
[named entity recognition] CoNLL-03 format, BIO/BILOU versions
[entity detection] ACE EDT guidelines (documentation)
[entity link tracking] ACE LNK guidelines (documentation)
[entity linking] TAC KBP EDL guidelines (documentation)
[relation extraction, event extraction] TAC Rich ERE guidelines (documentation)
[relation extraction] TACRED annotation scheme (documentation)
[entity tagging, values, relations, event extraction] ACE English guidelines for Entities, Values, Relations, Events

13.3 Annex Safety/Security

Table 17: Examples of existing tests and certifications for safety/security/ privacy

	Safety Product/System	Safety process	Security Product/System	Security process-oriented	Privacy Product/System	Privacy process-oriented
Goals of testing	Safety and health (Annex I Machinery Directive, particularly: Requirements for the control system (reliability (for criteria, see standards for the design of control systems (e.g. DIN EN ISO 13849-1:2016 [109] and machine type-specific Type C standards)), ergonomics if there is a human-machine interface). Reliability, Availability, Maintainability, Safety	Risk assessment and risk reduction (process according to DIN EN ISO 12100: 2011 [517] (machine type-specific type B and C standards)) Software – Development according to the V-model (DIN EN 61508:2011 [101], [102], [103], [433]) Consideration of hardware requirements	→ Confidentiality → Integrity → Accountability → Authenticity → Availabiltiy → Non-repudiation Security by design and default Security over LifeCycle	→ Confidentiality → Integrity → Accountability → Authenticity → Availabiltiy → Non-repudiation Information Security Management System (ISMS) including risk assessmen	**Data protection Privacy by design and default** Privacy impact assessment (impact/risk) **Data security** → Confidentiality → Integrity → Accountability → Authenticity → Availability → Non-repudiation	**Data protection processes / ISMS** Privacy impact assessment (impact/risk) **Data security** → Confidentiality → Integrity → Accountability → Authenticity → Availability → Non-repudiation
Types of assessment/tests Operationalization	Machinery Directive: Self-declaration EU type examination (1 type is examined – manufacturer ensures that production takes place in accordance with the type) CE-marking (CE = Conformité Européenne; European conformity) Self-declaratio	Machinery Directive: Self-declaration (Module A – internal production control (both product and process))	Self-declaration Marking	Certificate	Certificate	Certificate
Test provisions from	Machinery Directive: (Module A – internal production control (both product and process))	Machinery Directive: Quality assurance (Annex X Machinery Directive: Checking whether the development, manufacturing and testing process meets the requirements, criteria of module H)		e.g. ISO/IEC 27001 [480] ff. DIN EN IEC 62443 (all parts) [435]		DIN EN ISO /IEC 29100:2020 Privacy framework [133] ISO/IEC 27701 [128]

	Safety Product/System	Safety process	Security Product/System	Security process-oriented	Privacy Product/System	Privacy process-oriented
Certification to	Machinery Directive: machine type-specific type C standards harmonized standards → Annex I Machinery Directive	Machinery Directive: No standards exist that are considered sufficient on their own (e.g. DIN EN ISO 9001: 2015 [263])		NIS/CSA e.g. with ISO 27001 ff. or e.g. DIN EN IEC 62443 (all parts) [435]	GDPR certification in progress but not yet adopted.	GDPR
	Machinery Directive: GS mark ("Geprüfte Sicherheit" = safety tested) – national → Voluntary inclusion of a third party for machines)					

13.4 Annex Mobility

Trustworthiness-readiness matrices: selected results

Summary of relevancies and operationalization status and derivation of needs for action by means of the trustworthiness-readiness matrix (TRM)
In two workshops with experts for each of the three use cases

1. Evasive manoeuvres as a complex driving manoeuvre in automated driving
2. Ridesharing as a mobility service (mobility chain)
3. Traffic optimization via an improvement of the traffic signal control in the traffic infrastructure

in each case the

→ relevance and
→ operationalization status

for the two-dimensional space with the dimensions

→ embedding and life cycle phases and
→ trustworthiness aspects (hereinafter also referred to only as "TW Aspect")

were developed [312]. For this purpose, points were assigned to each cell of the matrix on a scale of 0 to 10 and colour-coded according to the significance (green = tends not to lead to a great need for action, yellow = possibly leads to a need for action, red = probably leads to a great need for action). From the resulting matrices for relevance and operationalization level, the following formula was applied

→ Need for action = 1,5 * Relevance * (10-operationalization) / 10

and the various needs for action were derived.

Finally, the exact point values were removed to prevent false accuracy or misinterpretation in this regard. Although the respective evaluations are based on the experience of selected experts, they were agreed upon by a small group of people. Consequently, there is a lack of strictly reliable and broadly agreed criteria for the respective point evaluations (these are currently being developed outside of the 2nd edition of the Standardization Roadmap).

In the following, the developed matrices are shown – as a supplement to the above summary in text form (cf. Chapter 4.6).

Use case evasive manoeuvres as complex driving manoeuvres in automated driving
Figure 54 shows the relevance of the combinations of TAI (trustworthy artificial intelligence) aspects and life cycle phases or embedding aspects for the use case evasive manoeuvres in automated driving.

Figure 55 shows the status of operationalization of the combinations of TAI aspects and life cycle phases or embedding aspects for the use case of evasive manoeuvres in automated driving.

Figure 56 shows the needs for action regarding the combinations of TAI aspects and life cycle phases or embedding aspects for the use case evasive manoeuvres in automated driving.

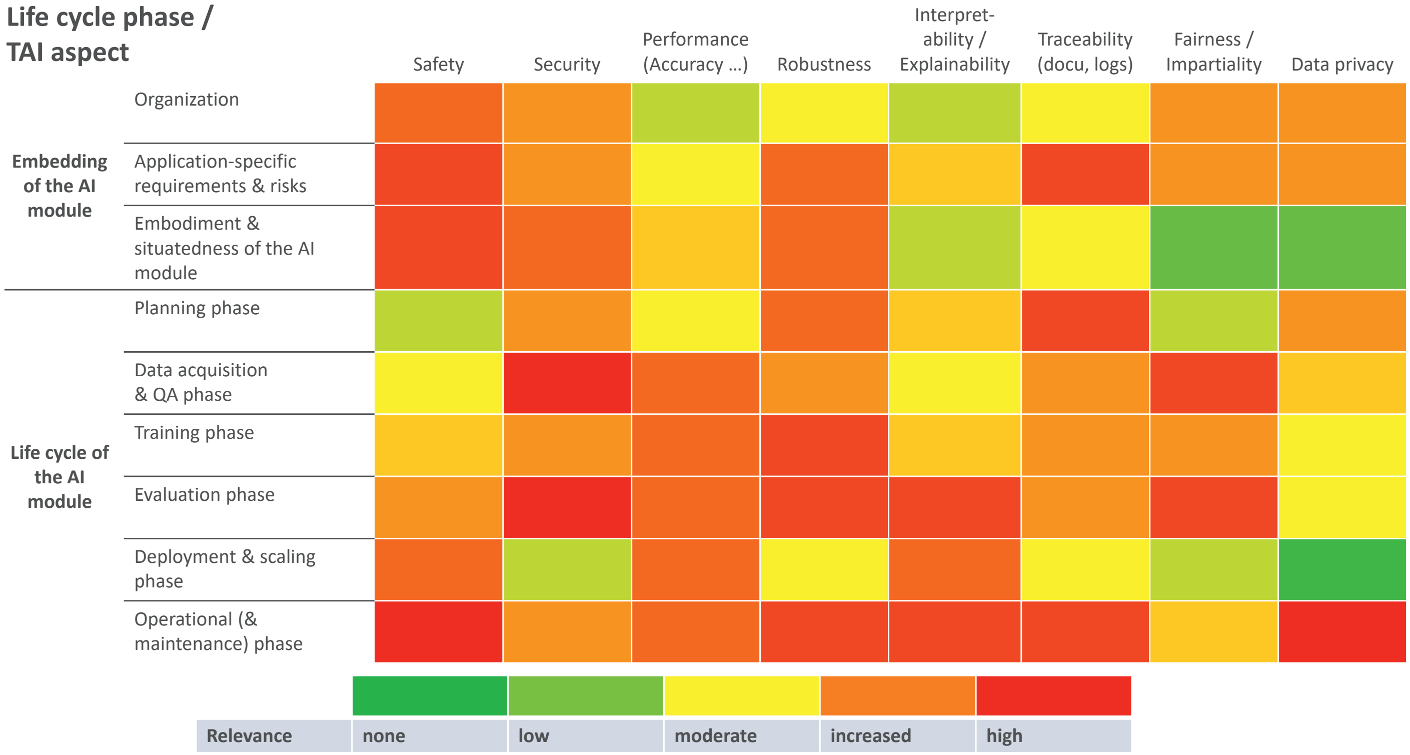


Figure 54: Relevancies of the combinations for the use case evasive manoeuvres (Source: Arndt von Twickel, Martin F. Köhler)

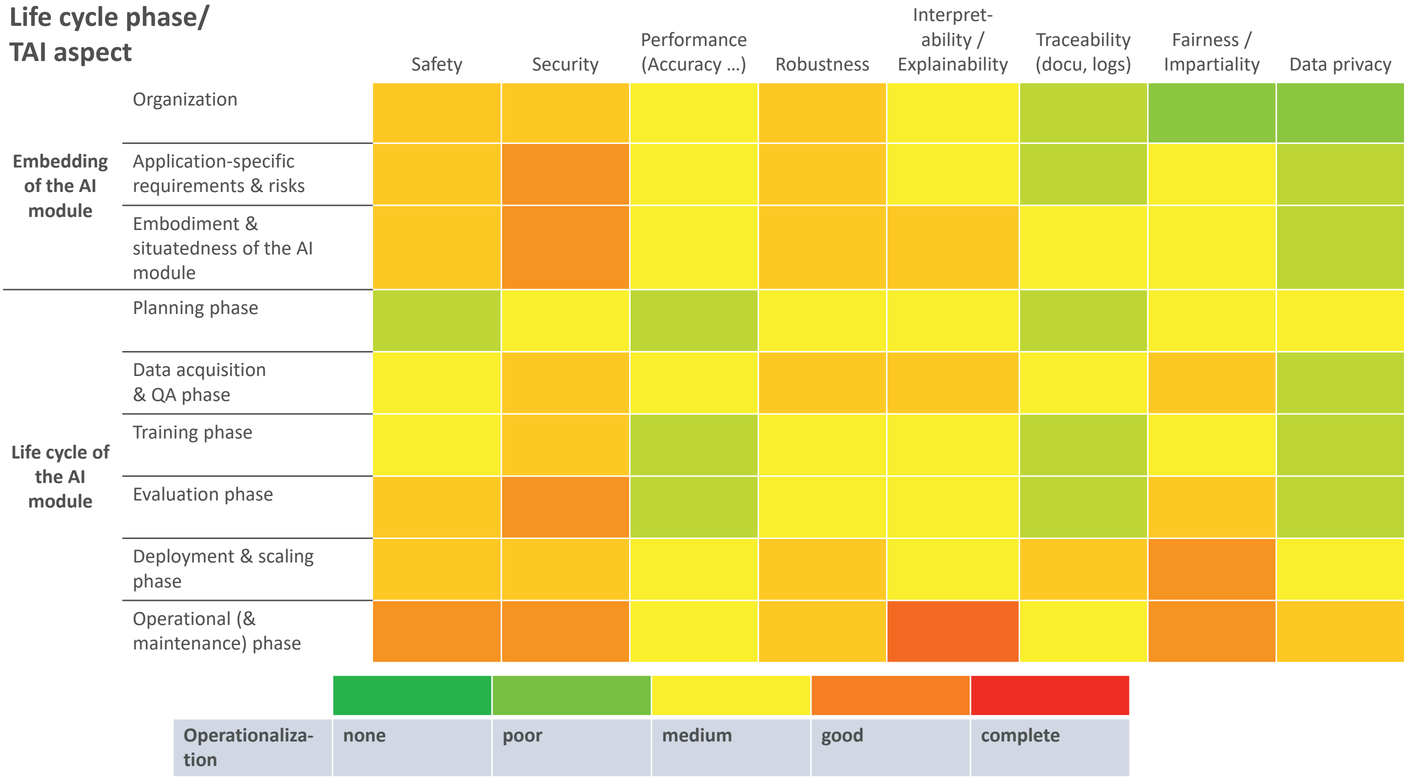


Figure 55: State of operationalization for the use case evasive manoeuvres (Source: Arndt von Twickel, Martin F. Köhler)

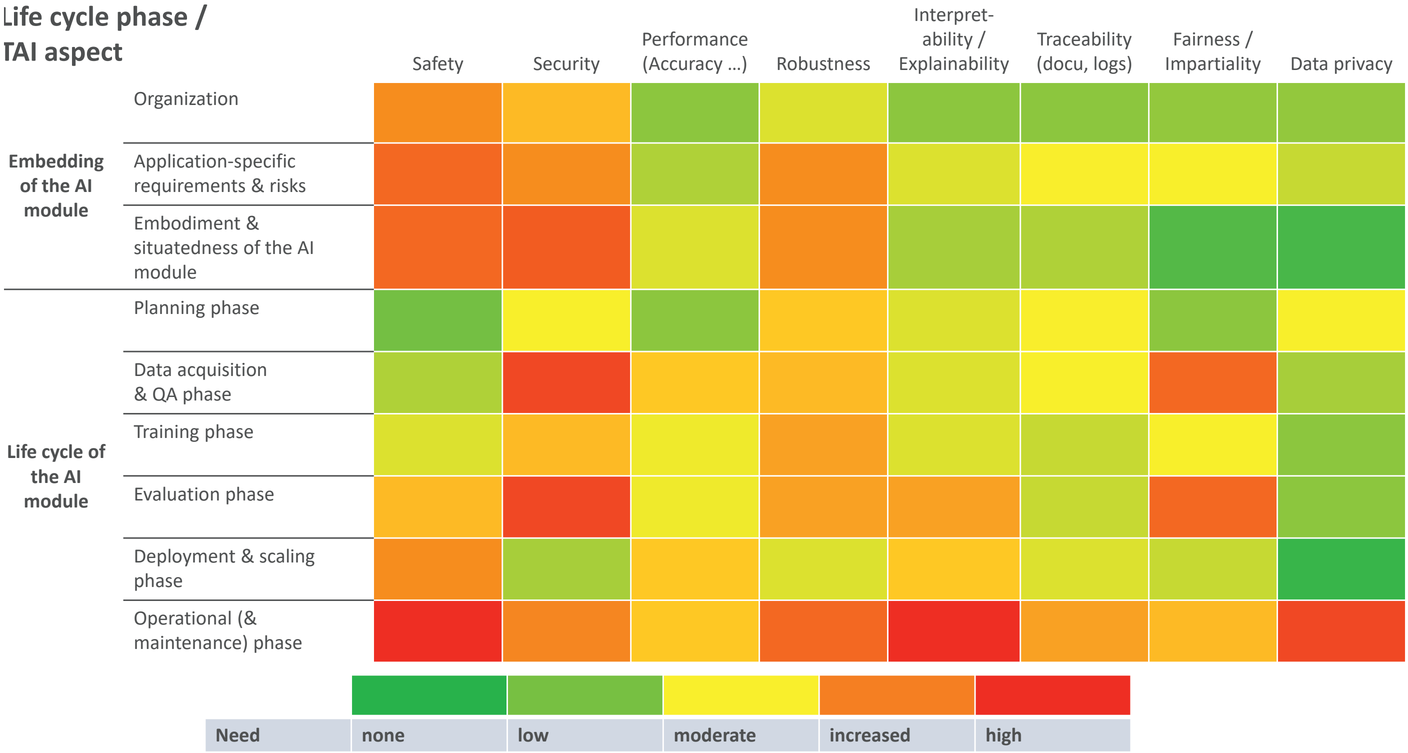


Figure 56: Needs for action for the use case evasive manoeuvres (Source: Arndt von Twickel, Martin F. Köhler)

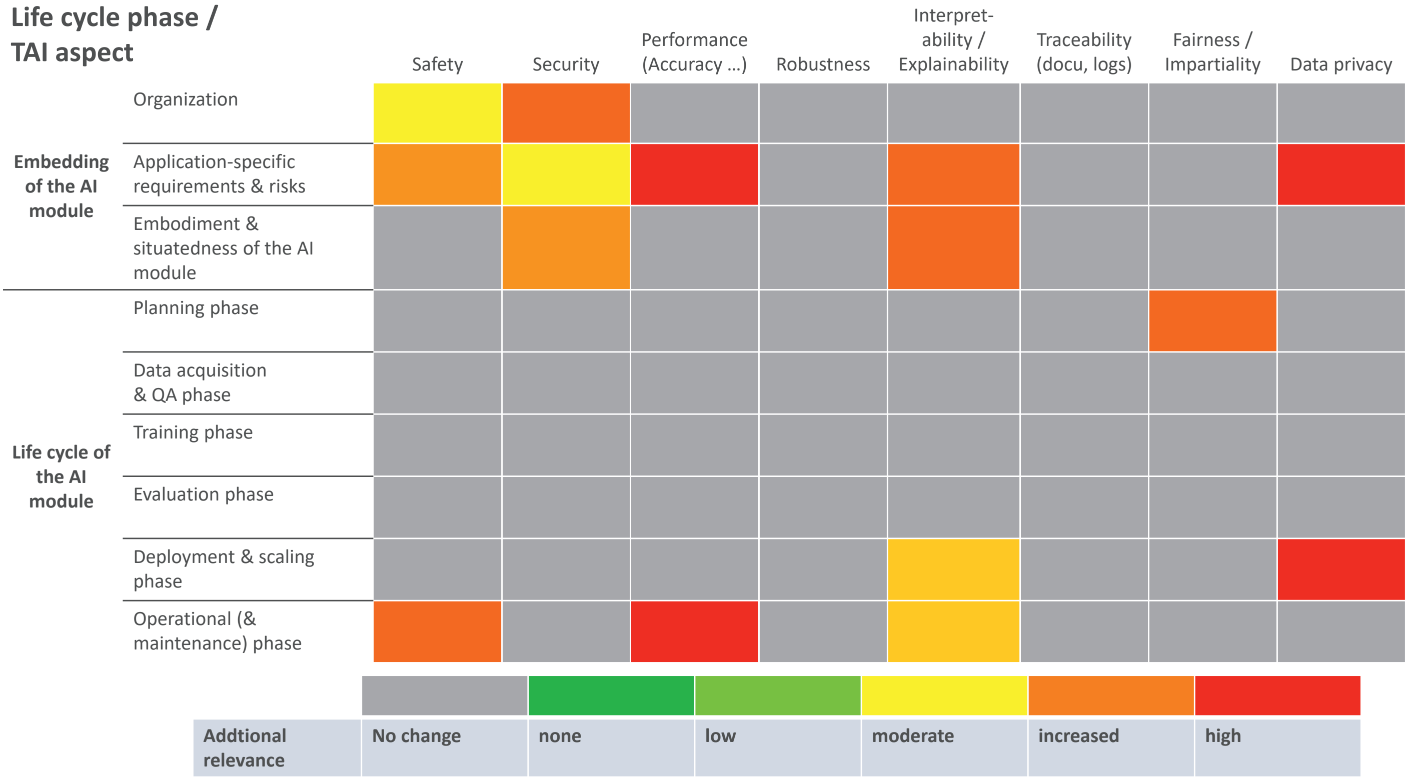
Use Case Ridesharing as a mobility service (mobility chain)

When recording the current trustworthiness readiness status for ridesharing, the current status in the area of automated driving serves as a basis and this is expanded to include ridesharing aspects. Therefore, only the additional relevancies, operationalization statuses and needs for action in relation to automated driving are listed here. The cells without significant changes are marked grey, those with changes in the same colour code as used above.

Figure 57 shows the relevancies of the combinations of TAI aspects and life cycle phases or embedding aspects for the use case ridesharing as a mobility service (mobility chain) – additions as compared to automated driving. Cells with unchanged relevancies are grey, the rest are marked in the familiar colour scheme (see above).

Figure 58 shows the status of operationalization of the combinations of TAI aspects and life cycle phases or embedding aspects for the use case ridesharing as a mobility service (mobility chain). The matrix was supplemented in comparison to automated driving. Cells with an unchanged operationalization status are grey, the rest are marked in the familiar colour scheme (see above).

Figure 59 shows the needs for action regarding the combinations of TAI aspects and life cycle phases or embedding aspects for the use case ridesharing as a mobility service (mobility chain). The matrix was supplemented in comparison to automated driving. Cells with unchanged needs are grey, the rest are marked in the familiar colour scheme (see above).



Life cycle phase / TAI aspect		Safety	Security	Performance (Accuracy ...)	Robustness	Interpret-ability / Explainability	Traceability (docu, logs)	Fairness / Impartiality	Data privacy
Embedding of the AI module	Organization	🟨	🟧	▨	▨	▨	▨	▨	▨
	Application-specific requirements & risks	🟧	🟨	🟥	▨	🟧	▨	▨	🟥
	Embodiment & situatedness of the AI module	▨	🟧	▨	▨	🟧	▨	▨	▨
Life cycle of the AI module	Planning phase	▨	▨	▨	▨	▨	▨	🟧	▨
	Data acquisition & QA phase	▨	▨	▨	▨	▨	▨	▨	▨
	Training phase	▨	▨	▨	▨	▨	▨	▨	▨
	Evaluation phase	▨	▨	▨	▨	▨	▨	▨	▨
	Deployment & scaling phase	▨	▨	▨	▨	🟨	▨	▨	🟥
	Operational (& maintenance) phase	🟧	▨	🟥	▨	🟨	▨	▨	▨

	▨	🟩	🟩	🟨	🟧	🟥
Addtional relevance	No change	none	low	moderate	increased	high

Figure 57: Relevancies of the combinations for the use case ridesharing (Source: Arndt von Twickel, Martin F. Köhler)

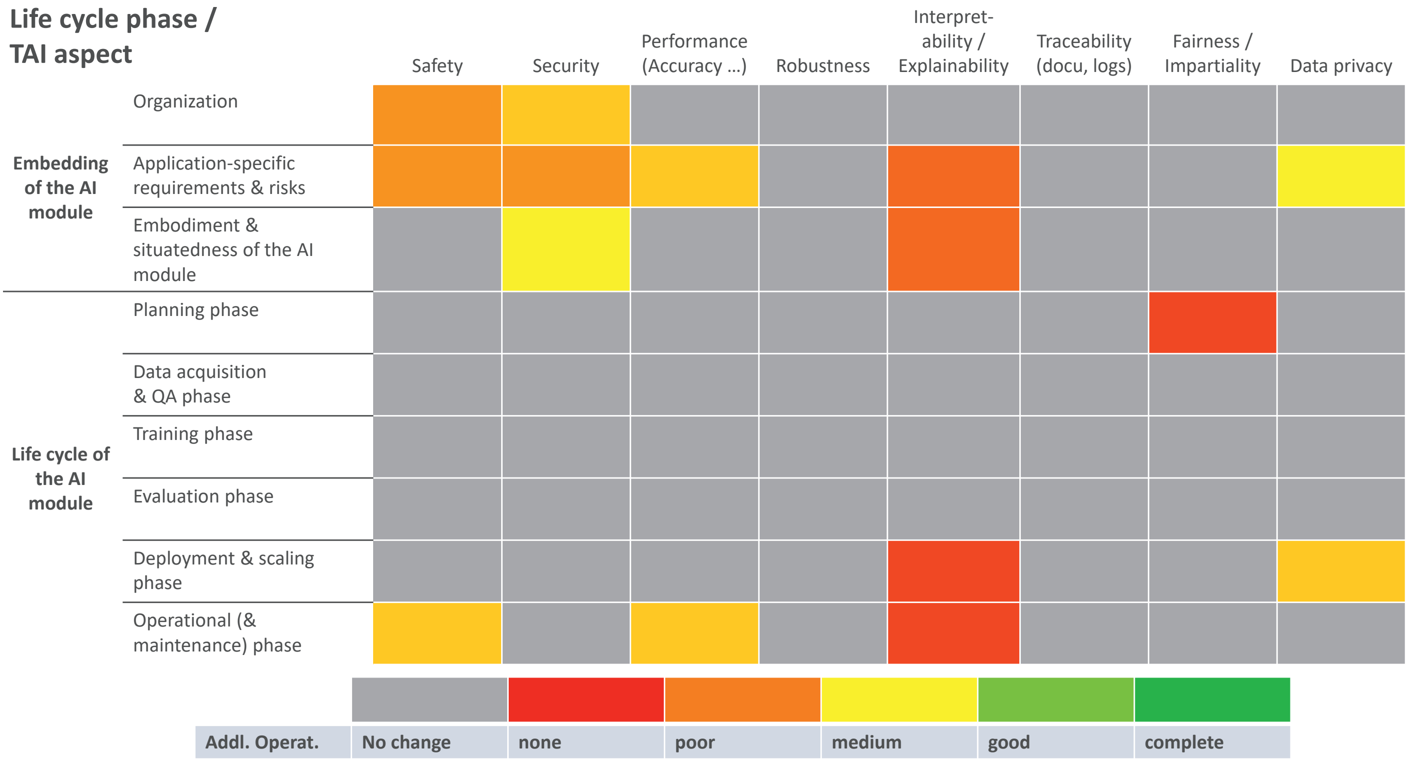


Figure 58: State of operationalization for the use case ridesharing (Source: Arndt von Twickel, Martin F. Köhler)

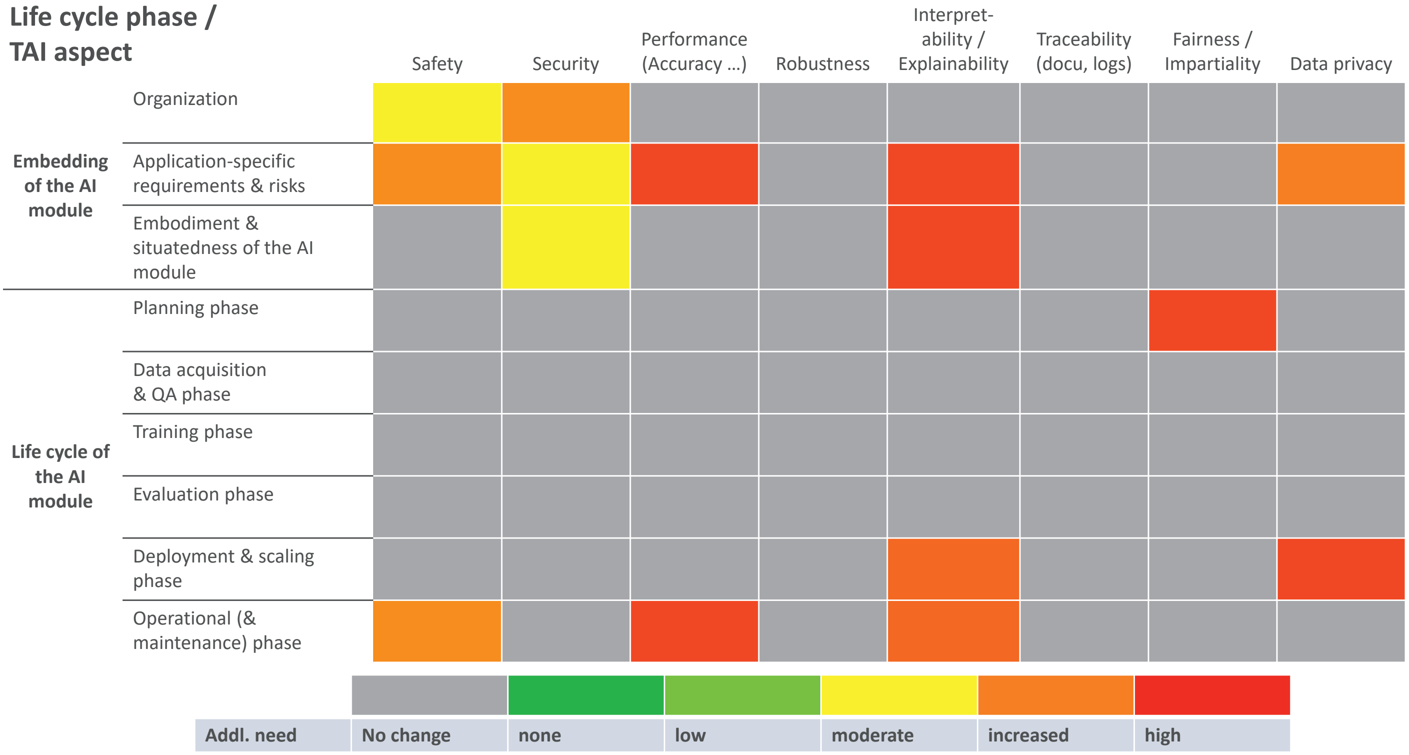
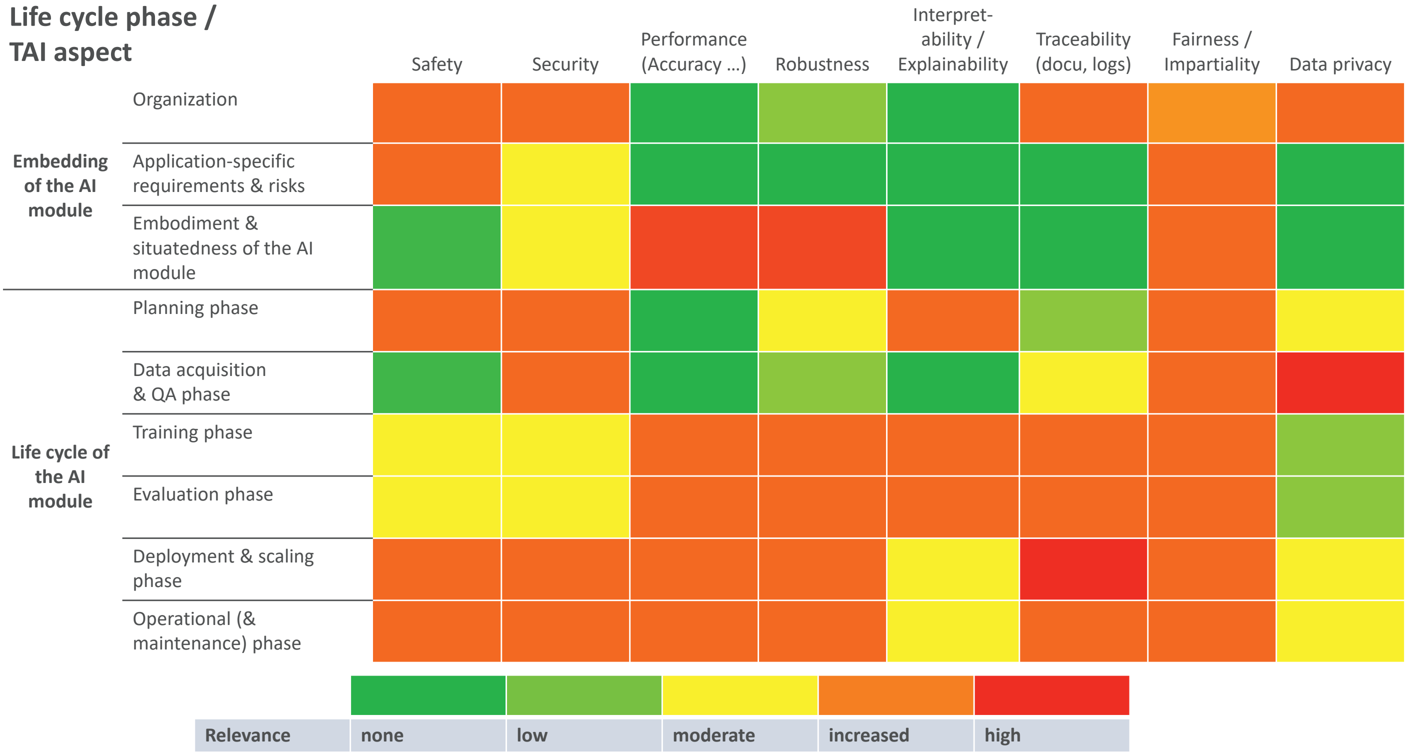


Figure 59: Needs for action for the use case ridesharing (Source: Arndt von Twickel, Martin F. Köhler)

Use case traffic optimization via an improvement of the traffic signal control in the traffic infrastructure
Figure 60 shows the relevancies of the combinations of TAI aspects and life cycle phases or embedding aspects for the use case traffic signal control in the transportation infrastructure.

Figure 61 shows the status of operationalization of the combinations of TAI aspects and life cycle phases or embedding aspects for the use case traffic signal control in the transportation infrastructure.

Figure 62 shows the need for action regarding the combinations of TAI aspects and life cycle phases or embedding aspects for the use case traffic signal control in transportation infrastructure.



Life cycle phase / TAI aspect		Safety	Security	Performance (Accuracy ...)	Robustness	Interpret-ability / Explainability	Traceability (docu, logs)	Fairness / Impartiality	Data privacy
Embedding of the AI module	Organization	🟧	🟧	🟩	🟢	🟩	🟧	🟧	🟧
	Application-specific requirements & risks	🟧	🟨	🟩	🟩	🟩	🟩	🟧	🟩
	Embodiment & situatedness of the AI module	🟩	🟨	🟥	🟥	🟩	🟩	🟧	🟩
Life cycle of the AI module	Planning phase	🟧	🟧	🟩	🟨	🟧	🟢	🟧	🟨
	Data acquisition & QA phase	🟩	🟧	🟩	🟢	🟩	🟨	🟧	🟥
	Training phase	🟨	🟨	🟧	🟧	🟧	🟧	🟧	🟢
	Evaluation phase	🟨	🟨	🟧	🟧	🟧	🟧	🟧	🟢
	Deployment & scaling phase	🟧	🟧	🟧	🟧	🟨	🟥	🟧	🟨
	Operational (& maintenance) phase	🟧	🟧	🟧	🟧	🟨	🟧	🟧	🟨

	🟩	🟢	🟨	🟧	🟥
Relevance	**none**	**low**	**moderate**	**increased**	**high**

Figure 60: Relevancies of the combinations for the use case traffic signal control (Source: Arndt von Twickel, Martin F. Köhler)

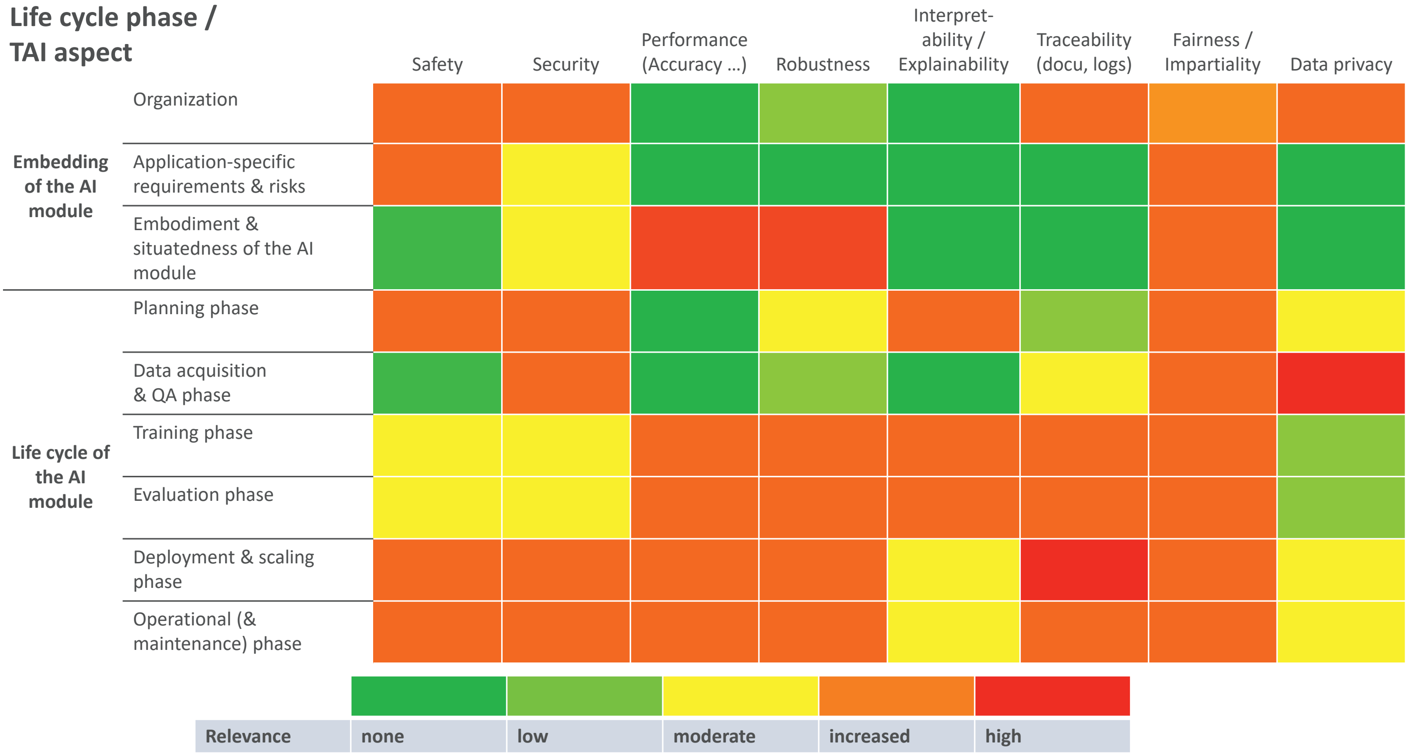


Figure 61: State of operationalization for the use case traffic signal control (Source: Arndt von Twickel, Martin F. Köhler)

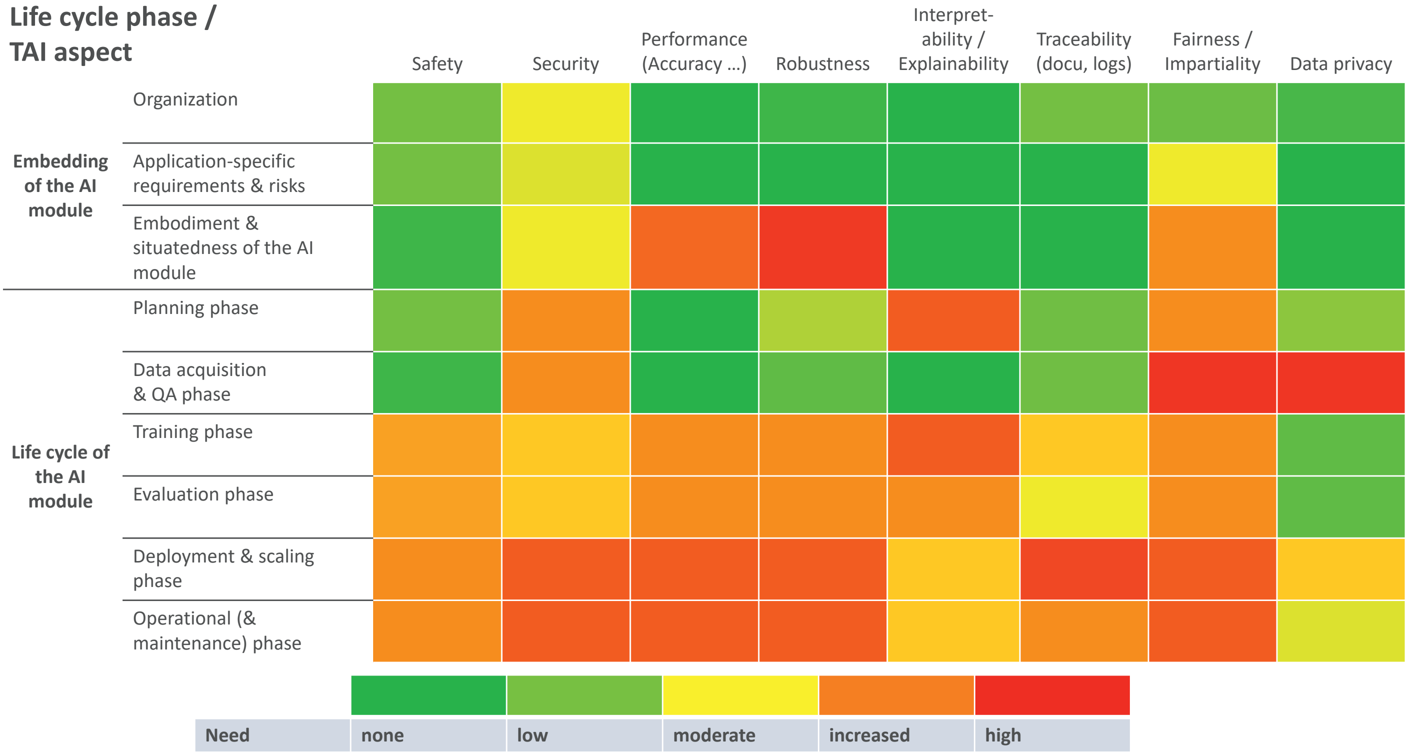


Figure 62: Needs for action for the use case traffic signal control (Source: Arndt von Twickel, Martin F. Köhler)

13.5 Annex Medicine

To provide an overview of the application examples of Chapters 4.7.2.1 to 4.7.2.3, the three AI-based medical applications were examined according to the following comparative criteria:

→ Actors (persons involved)
→ Goal (description of the problem solved by the medical device)
→ System (description of the mode of action of the medical device)
→ Prerequisite (technical, organizational or infrastructural requirements for service provision)
→ Trigger (what triggers the application?)
→ Stakeholders (other parties interested in the AI application)

Table 18: Use case 1: AI-assisted 2-D X-ray image analysis for caries diagnostics in dentistry

Actors	→ Attending dentist → Specialist (e.g. diagnostic radiology, oral and maxillofacial surgery), where applicable → Doctor making the referral, where applicable → Medical-technical assistant → Patient → Developer → Health system is indirectly involved → Health insurers are indirectly involved
Goal	Assisted 2-D X-ray diagnostics with the aim of saving time and enhancing the reproducibility of diagnostic workflows. Where appropriate, improved diagnostics and therapy options through AI-supported 2-D X-ray image analysis (benefit for all actors involved).
System	The system consists primarily of a software component that detects anatomical and, where appropriate, pathological features in 2-D X-ray images and visualizes them for dentists, i.e. marks them on displayed images. Such a component can, for example, be integrated as a backend service in a larger software architecture. The input data for this component consists of a 2-D X-ray image and meta information (e.g. pixel size, radiation dose); the component provides contours (2-D polylines) and annotations for each contour (depending on the object, e.g. tooth number). The execution of the component is initiated by the radiological system (calculation after availability of a new 2-D X-ray image), and the numerical results are stored in a database together with the patient and image data. Visual representation of these results is provided by the attending medical staff at a workstation connected to the system after the dataset is opened. The attending physician or specialist examines the displayed result and makes manual corrections if necessary. Such corrections are transferred to the above-mentioned database. The downstream diagnosis is performed according to dental standards on the basis of the image data and with the aid of the (corrected, if necessary) AI-assisted information.
Prerequisites	Practice qualified for dental X-ray diagnostics with appropriate staff and technical equipment/ infrastructure This includes: → X-ray device → Medical-technical assistant for taking the X-ray image → Software with “AI components” → Physician who is familiar with the system for the evaluation of the 2-D X-ray images

Trigger	The component is called automatically after the system provides a new dataset.
Stakeholders	→ Data Protection Officer: The component modifies the patient dataset; security of data transferred and stored in the database must comply with applicable privacy policies. → Developer → Hospital with IT department → Health insurers → Notified bodies regarding the implementation of conformity assessment

Table 19: Use case 2: Ventilator with AI-assisted weaning

Actors	→ Patient (here: person, pediatric 15-35 kg, adult 35-200 kg) → Specialist in anesthesiology/intensive care medicine → Intensive care nurse → Medical technician → Manufacturer
Goal	Use case: Ventilator with AI-assisted weaning The Corona pandemic in particular has shown that gentle weaning adapted to the patient’s needs is crucial for rehabilitation and sustained well-being. Another effect is the reduction of the workload in intensive care, because in the usual clinical procedure, the ventilation parameters have to be adjusted manually again and again during weaning, depending on the patient’s condition. With an automated system, appropriate adjustments are made at shorter intervals, thus shortening the overall weaning process and providing better situational support for the patient. The number of near-body contacts of ICU nurses with potentially infectious ICU patients is also reduced. Compared to the pre-existing automated system based on classical AI, the new neural network-based system offers the advantage of being able to learn from intensive care physicians and nurses to ensure an appropriate response to more and more exceptional situations, thus reducing the current flood of alarms to the really important alarm situations.
System	**1. System description** The weaning system is integrated in the ventilator and realized as a neural network (NN). This is a “frozen” NN, meaning that the learning phase is completed before the device is released to the market. The task of the weaning system is to support the adaptation of intubated patients during weaning from the ventilator. Patients are already able to initiate spontaneous breaths, but not forcefully enough to fight the resistance of the tube and get enough air. Therefore, they are supported with positive airway pressure. Support should be gradually reduced to return patients to normal breathing. However, if respiratory deterioration or stress symptoms occur, support must be increased again. The weaning system must be able to correctly classify the patient’s condition in terms of ventilation needs (diagnostic function) and adapt the respiratory support accordingly (therapeutic function). The NN comprises three layers: 1) Input layer, 2) Diagnosis layer, 3) Output layer.

In the input layer, there are nodes for parameters that are set at the beginning of weaning, such as patient class (pediatric/adult), weight, height, medical history (e.g., for patients with COPD or neurological disorder). Three input nodes are fed with parameters continuously measured by the ventilator: fspn (frequency of spontaneous breathing), VT (tidal volume), $etCO_2$ (end tidal concentration of carbon dioxide).

There are eight nodes in the diagnosis layer, each of which correlates to a classification of the patient's condition with respect to breathing. The eight states are: normal ventilation, hyperventilation, tachypnea, severe tachypnea, insufficient ventilation, hypoventilation, central hypoventilation, unexplained hyperventilation.

In the output layer, the therapeutic decision is made based on the diagnosis. Here there are three nodes: a) decrease of pressure support, b) increase of pressure support, or c) alarm without change of pressure support.

After an alarm is raised, the system stores the extent of pressure correction required by hospital staff. The state of all parameters when the alarm goes off and after correction is also saved. These data are made available to the manufacturer via a data interface, either directly or indirectly through hospital staff.

The parameter sets collected from the devices in the field during an alarm situation are used to teach or test a new, improved network NN2 at the manufacturer. The NN2 is then analyzed by the manufacturer, compared to the NN, and if found to be more appropriate after thorough benefit-risk assessment, frozen and used for a new version of the system if necessary. Objectives of the change may be: Shortening of weaning, less and lower fluctuations in performance, reduction of situations leading to an alarm, better adaptation to unusual situations, elimination of detected bias.

2. Performance process

A long-term ventilated patient is stable enough for the specialist to order the start of weaning. Automatic weaning is started manually (doctor/nurse on doctor's orders).

The system enters the "adjustment" phase. Here, the NN is used to make periodic downward (a) or situational upward (b) adjustments to airway pressure support.

Only in the case of output c) (see system description) is an alarm raised, and then the intensive care nurse intervenes. The relevant parameters and the pressure correction made manually are stored in the system and can be forwarded anonymously to the manufacturer for optimization of the NN.

If the pressure support falls below a certain threshold required for tube compensation, the system enters the "observation" phase, which lasts one to two hours depending on the initial level of pressure support. This phase corresponds to an automated spontaneous breathing test. The downward adjustment of the pressure support must not fall further below the threshold value. The system reports a successful weaning of patients when respiratory instabilities remain below 20 % of the observation time, i.e. below 12 to 24 minutes. Otherwise, the system goes back to the "adjustment" phase.

If the "observation" phase is successful, the system switches to the "maintenance" phase. Patients continue to be ventilated with constant low pressure support, with minor instabilities being compensated for as in the "observation" phase. Only in the event of frequent or prolonged instability is the weaning message withdrawn and the system automatically returned to the "adjustment" phase. For reasons of transparency and situational awareness, this should also be reported/displayed. During the "maintenance" phase, it is recommended that the physician order extubation at any time.

Prerequisites	→ Data port on the ventilator and internet connection of the hospital, in order to be able to transmit relevant parameters of the system to the manufacturer, as in alarm situations, and to save them as datasets for future learning or test phases, as well as for the evaluation of possible undesired system behaviour (bias). → Willingness of the hospital to make the anonymized usage data and parameters available to the manufacturer for further development of the system. → If the anonymization of the data is not sufficiently possible, the consent of the patients or their relatives for the use of the data must also be obtained and the manufacturer must ensure the confidentiality of the data that cannot be anonymized.
Trigger	→ Medical decision to start weaning (not automatic).
Stakeholders	→ Regulators → Data protection officer → IT director of hospital → Owner of/body responsible for hospital → Health insurers

Table 20: Use case 3: Segmentation and classification of brain areas (including cerebrospinal fluid) and their volume determination

Actors	Direct: Physicians (from the fields of radiology, neurosurgery, neurology) (trigger analysis, evaluate results, make decisions, prognoses, diagnoses based on these) Indirect: Patient
Goal	The chosen application example solves the problem of time-consuming, manual or simple partially assisted segmentation of structures in the image. Originally very time-consuming manual work is automated and performed in a clinical context with higher accuracy and repeatability. This concerns in particular occurring inter- and intra-individual deviations in repetitions. It is used by physicians to support diagnoses of neurodegenerative diseases.
System	The application example describes an AI-supported, fully automated segmentation of all relevant brain areas based on 3D MRI (magnetic resonance imaging) data. The segmented regions are volumetrically quantified and visualized. The calculation is performed by receiving the data from the radiological infrastructure (Picture Archiving and Communication System, PACS).
Prerequisites	Technical: Server infrastructure, connection to imaging workstation and/or PACS Suitable devices for generating 3-D imaging data (such as the 1,5 Tesla (T) MRI scanner).
Trigger	The trigger for segmentation and volume determination is the transmission of the image data to the radiological infrastructure (PACS).
Stakeholders	→ Manufacturer → Notified body (conformity assessment) → Supervisory bodies → IT department (technical implementation) → Data protection officer (transfer of data to the system) → Health insurers (possibly re the early detection of diseases).

13.6 Annex Energy/environment

To provide an overview of the use cases in Chapters 4.9.2.1 to 4.9.2.6, the use cases were systematically analyzed and structured as follows:

→ Actors (persons involved)
→ Goal (description of the problem to be solved)
→ System (description of the mode of action)
→ Prerequisite (technical, organizational or infrastructural requirements for service provision)
→ Trigger (what triggers the application?)
→ Stakeholders (other parties interested in the AI application)

Table 21: Use case 1: Autonomous Smart Grid Power Management and Consumption System

Actors	→ Electrical Power Management System (PMS) Energy Provider → Electrical System Interface (SIF) Manager → Distributed Energy Resource (DER) Manager → Industrial Automation and Control System (IACS) Energy Consumer → Layered Communication IT Operator (Communication) → Value Stream Life Cycle Operator (Semantics) → AAS Asset Operator (Physics) → Digital Twin Operator (Analytics, causes) → Data Manager (Learning, effectiveness)
Goal	→ Production and products that can be adapted (parameterized) to customer wishes or ethical requirements (asset/value stream operator roles) → Flexible (smart) generation, transmission, distribution, and consumption of energy (PMS/IACS/SIF/DER manager roles) → Collecting, presenting and acquiring knowledge (analyst role) → Structuring asset data spaces (data management role)
System	Industrial reference models such as SGAM or RAMI4.0 describe the structure of systems-of-systems. The system structure includes a) the ontologies of structural (syntactic) interoperability, b) the semantic interoperability in the value stream between semantic domains (called conduits) during the life cycle, and c) the physical hierarchies, i.e., usage structure (called zones) of the asset under consideration and its AAS. The power in SGAM systems is the effective energy supply of industrial production plants (according to RAMI4.0) or individual consumers. From generation to consumer, energy in the value stream (SGAM x-axis) takes various heterogeneous forms depending on the medium it must carry. These are also referred to as heterogeneous models, which must interact semantically. These energy sources can be weather patterns if wind and solar are viable volatile sources. Energy carrier or energy generating media are mechanical wind generators or photovoltaic devices, and long-range high-voltage DC transmission grids or local AC grids are required for electrical power transmission. Distributed energy resources (DER) are used to distribute energy in the SGAM network. And finally, consumers are dependent on its production facilities, on its energy consumption behaviour, on the availability of electrical energy and, last but not least, on the “ethical quality” of the traded energy.

	The performance in RAMI4.0 systems is the efficient production of a product. In the value stream (x-axis RAMI4.0) for manufacturing a product, the product, similar to energy, takes on different heterogeneous formats depending on its life cycle state. These formats differ roughly in typification and instantiation of the product. Both typification and instantiation are characterized by a development and usage phase These sequential production states can be modelled with different models that must interact semantically. So, from a semantic point of view, there is definitely a comparability of the value stream in the reference architecture models when modelling the states and transformation of the properties of product development or energy supply.
Prerequisites	In order to function autonomously (or automatically) to a large extent, both production plants and energy supply systems require infrastructure measures that enable them to receive and process information from the environment (from the outside) as well as from the embedding (from the inside). While the external environment is unknown and therefore needs to be learned, the embedding is known a priori and can be fed into the value stream as a model. Known process models can be automated, and unknown process models can be learned using appropriate ML methods and used to autonomize energy transmission or production. Automation and autonomization can both be combined for system control. In this context, automation represents a closed loop and autonomization an open loop. The adjectives “open” and “closed” denote circuits that are open or closed to the outside world as well as to their environment with respect to the reception of information.
Trigger	In SGAM systems, the triggering function is, for example, the unexpected availability of volatile energy due to weather patterns. Before connecting the additional power, the network managers must simultaneously measure, analyze and decide on the stability of the electrical supply network under the given and the changed conditions. This decision has implications for the downstream networks or “energy carrier media” to generation, transmission, distribution, consumption. All networks or media must be harmonized, i.e. coordinated with each other, to avoid instabilities.
Stakeholders	Automated or autonomous control of energy grids as critical infrastructures of hospitals, public safety, traffic control, weather forecasting, etc. provides a host of other dependencies of critical infrastructure stakeholders on reliable and ethical energy supply.

Table 22: Use case 2: Energy efficiency in buildings and coupling with energy networks

Actors	→ Building operators → AI developers
Goal	→ With greater use of renewables, electricity production fluctuates with the weather. → Energy must be consumed more when it is generated. → Buildings are to be used as flexible energy consumers in the power grid. → Energy use in the building should be optimized (40 % of energy is used in buildings). → Intelligent control of air conditioning, heating, water heating and charging stations.

System	The AI system will be connected to building controls as well as to weather and energy grid data. Using this data, it generates a daily forecast of building usage as well as renewable energy availability. An intelligent algorithm then controls the building to use energy mainly during periods of high renewable availability. **Components:** In the building → Temperature, humidity, CO_2 sensors → Building management system → Control units of air conditioning, heating, charging stations, etc. Cloud/Internet → Dashboard app: displays sensor values, takes user input to set AI system boundaries → AI backend: determines forecasts and control commands → Energy grid data interface → Weather data interface **Process:** Every day at midnight: → Retrieve weather report → Retrieve power grid forecast about availability of renewables → Perform building occupancy forecasting (based on historical occupancy data) → Simulation and optimization of building energy use → Create an optimized 24-h schedule for flexible device control During the day: → Control of the building based on the created plan → Adaptation of the plan to real-time changes
Prerequisites	→ Integration of the software into the building, creation of interfaces to the building management system → Integration of energy grid and weather data → Training the ML models to the building data → Setting up the software on a cloud system
Trigger	→ Automatic trigger every day at midnight to create a new plan
Stakeholders	→ Energy network operator → Data protection officer (prediction of room occupancy partly critical)

Table 23: Use case 3: Personalized AI-powered recommendation systems for sustainable consumption

Actors	End consumers, trade and any actors in the value and supply chain, data hub to the DPP and operators of the AI
Goal	→ **Problem:** The lack of transparency and clarity of product-related sustainability information when purchasing (e.g. „label jungle“) is an obstacle to sustainable consumption ([418], [419]). As a result, consumers select products that do not match their individual attitudes, including sustainability preferences. This problem affects both stationary and online retail. → **Solution part a) Personalized recommendation systems for sustainable consumption through AI-supported assistance systems for concrete purchasing decisions** ([424], 54) (based, among other things, on environmentally related life cycle data, possibly as data from the DPP and personal preferences). This can increase the personal relevance of products in terms of **„meaningful product advice“** ([420], 12), as AI could match personal attitudes, preferences, and needs with product features ([424], 255, [524], 18). A comparison of individual sustainability preferences with a product database and personalized product recommendations generated from this enables, among other things, a sustainability optimization of the product selection. → **Solution part b) Strategic recommendations in consumption/purchase planning to optimize consumption patterns in line with demand** in the medium to long term (building on a)). By collecting and interpreting personal purchasing data, the recommendation system could gain important **insights into consumer behaviour.** AI could forecast product demands, flag frequent users, and suggest derived relevant alternatives. ● Example: If consumer A buys two packets of butter per week, an alternative suggestion of the recommendation system could be to choose the bulk pack for future purchases (less packaging material, cheaper, reduction of purchases). → In addition, supply chain actors can feed further information (e.g., nutritional information, recommended daily allowances, shelf life) into the recommendation system, enabling **optimization of consumption patterns** ([420], 24). Example: Consumer A from the previous example consumes comparatively more CO_2-intensive products and would also benefit health-wise from a less fatty diet. ● Example: The AI-assisted recommendation system might suggest choosing a large pack of vegetable margarine now instead of the two small packs of butter. Altogether, consumption patterns can be influenced and sustainably changed.
System	Technical system: → Emissions, resources – Prediction: uses regressive neural networks to determine the consumption of resources and emissions for a product. → Recommendation system: recommends sustainable products to users based on their and others‘ behaviour Personalized AI-powered recommendation systems could operate according to the logic of the following behaviour change mechanisms, among others: → **Coercive intervention/choice restriction:** Exclusion of non-sustainable options by AI with the disadvantage that no self-reflection of consumers is supported ([423], 12) → **Persuasive intervention/nudging:** Incentives to choose sustainable products/services within existing choices [(Thorun et al. (2017), 48f.)], although the moral power to decide which products to recommend rests with the designers/researchers ([423], 12). The Green Consumption Assistant project uses scraping technologies to develop a database of sustainable products [418]. **Reflected intervention/feedback:** Sustainable consumption decisions are made based on fed-back data/feedback (e.g., CO_2-emissions from last purchase) ([423], 12).

Prerequisites	**Technical requirements:** Data availability, standardized interfaces and structured data formats, development of algorithms, cloud infrastructure for execution, app to display recommendations **Organizational requirements:** Briefing customers on how to use system, briefing supply chain actors on how to feed information into system, clarifying which institution is running the AI. **Data basis:** 1. product-related (DPP, if applicable): sustainability-related characteristics (regionality, labels, obsolescence data, fishing methods, ingredients, allergens, parent companies) ([420], 20). 2. consumer-related: clustering of consumers (as AI training basis), purchasing behaviour, lifestyle, dietary habits
Trigger	Triggers for the AI system can be, for example: → Scanning / purchasing of item → Daily retraining of the algorithms, based on new data
Stakeholders	Use case actors, health insurers, research institutions, environment, consumer protection, data protection officers

Table 24: Use case 4: Scalable determination of environmental impacts in the building sector

Actors	→ Users ● Building and neighbourhood planning ● Formulation of political framework conditions for structural support measures → Server and DB maintainer (database)
Goal	The ecological life cycle analysis of buildings and neighbourhoods requires a great breadth and depth of information. This implies a great amount of time and computational effort in the determination of environmental impacts (cf. [425]). Machine learning methods can be used to determine and utilize guideline values that provide sufficient information about the footprint of the building/neighbourhood, as well as possible environmental optimizations. This results in significant time savings.
System	→ Frontend ● For user input (users) of key data on building/neighbourhood, which are basically available/publicly available in early planning phases and have highest entropy/information density in ML model ● For the display of the determined footprint and the statistical uncertainty → Backend ● For the determination of the footprint on the basis of the transferred user input by means of comparison with DB (model output) ● For anonymization of input data at a high level of detail and integration as training data into the ML model ● System component for the determination of learning/scale effects in the „Goal" ● System component to track and, if necessary, conformity check the data usage period ● Uniform data systematics (ontology, semantics)

Prerequisites	→ Technical/infrastructural ● Server(s) with pre-trained ML model and DB for model input/output ● REST-API (data interface) • For user input queries • Possible connection to ML server • Playing back the appropriate output from DB to frontend • Frontend as web service/plug-in/... → Organizational ● Process-related integration of the use case into the planning process of buildings/neighbourhoods or new buildings/refurbishments
Trigger	→ Input/upload of key data/files in frontend
Stakeholders	→ Data protection officer regarding anonymization and aggregation level of ML input and output → Data science actors → Simulation scientists

Table 25: Use case 5: Resource intensity of AI & ML

Actors	→ Users of systems with AI elements
Goal	Artificial intelligence and machine learning models, by definition, require a significant amount of data (processing) for pattern recognition. This tends to result in a high computing time and performance. This results in high energy consumption and associated environmental impacts (cf. [398]). For the intended application of AI elements, a meta-system should therefore be used to check whether the designed AI system/ML model has a need for optimization with regard to the required amount of data and algorithms and/or runtime.
System	→ Frontend ● For the user input (user) about the intended system (data set//algorithm//technical setup, ML server//...) ● For feedback • Predicted environmental impact of the system/AI element in the system • Optimization proposals for systematics based on ensemble learning results/confirmed research results → Backend ● Ensemble learning server ● Matching the input parameters with DB • For the environmental impact forecast • For finding optimization proposals based on previous ensemble learning/validated research results

Prerequisites	→ Technical/infrastructural ● (Server for) DB with categorized research results/comparative values, environmental indicators for calculating environmental impact ● Frontend as web service/plug-in/... ● REST-API (data interface) • For user input queries • Playing back the output from the backend to the frontend • Possible connection to ensemble learning server → Organizational ● Process-related integration of the use cases in the conception of AI applications ● Increased attention to resource intensity of AI applications ● Establishment of a uniform metric and/or reference system for comparability
Trigger	→ Input/upload of user input into frontend (manual/automated)
Stakeholders	→ Data protection officer → Data science actors

Table 26: Use case 6: Adversarial resilience learning – Market intervention by aggregators in the distribution grid

Actors	→ Distribution system operator: provides a local market to resolve grid congestion. → Consumer/prosumer: (regular) participants in the local energy market → “Attacker”: consumers/prosumers who want to use market rules to their advantage; concentrated on one line to increase effectiveness (if necessary also AI-based as „automatic attacker“ on the market) → “Defender”: Learning agent system for detection of market-based attacks
Goal	In the context of distribution system operation, market conditions may cause collusion and coordination of assets to be intertwined with knowledge of the state of the grid in such a way that gamification may occur: The aggregators optimize their coordinated behaviour in such a way that they create a bottleneck vis-à-vis the grid operators, which they can also unblock themselves, of course in return for appropriate compensation from third parties as an incentive. The market rules are thus “exploited” in such a way that artificial problems are triggered by the incentive of remuneration and then eliminated themselves in return for “remuneration”. Supply and demand are artificially created here. Goal: AI as attacker (market participants with malicious intent behaviour) simulates the gamification of the market, reveals vulnerabilities, and helps determine risk. Goal: AI as defender learns to detect the behavioural patterns of aggregators/malicious market participants and can thus (1) either act as an assistant system/detector or (2) initiate actions to prevent gamification (as a direct actor clearing house; depending on the market design).

System	Grid congestion is now a permanent problem for DSOs (distribution system operators). Redispatch processes are increasingly relying on small(est) plants (currently from 100 kW), while hardly any distribution grid is sufficiently equipped with sensors and actuators, or the ICT of the DSOs is not equipped to achieve a centrally controlled resolution of this increasingly complex situation. As an alternative, local markets are increasingly being considered or even implemented, which address the complexity problem through a form of self-organization. In the simple market design, the DSO determines a congestion situation on one line (radial feeders, not fully meshed network). It uses the flexibility market to encourage local consumers to reduce/shift load through financial incentives. In one variant of the use case, the goal can be to make the line as self-sufficient as possible, so that prosumers are encouraged to feed in accordingly via price signals. The grid operator typically cannot detect bilateral agreements between local participants. Load ramps can be artificially triggered/forced by participants, for example, by charging “electric vehicles (EV)” at specific times. Thus, the bottleneck is detected, but the trigger is officially unknown to the DSO. However, since each market participant can of course reduce the charging of its EV (or its load demand in general) in the same way, a coalition of market participants can draw money from the DSO almost at will without having to make a real counter-performance as a “sacrifice” in the form of a “real” incentivized change in behaviour. Since the coalitions are usually formed dynamically and the changes in the marginal distributions are not detectable, especially because the reason is unknown – of course, there are just as many valid, context-free reasons – this attack at the expense of the DSO is not detectable with previous means in the control technology.
Prerequisites	Collection of measurement data, scenario definition, algorithms about the market processes, topology data
Trigger	Congestion in the grid occurs unexpectedly and is suddenly resolved against incentive after price signals in the market. Frequent suspected gamification, AI responds autonomously based on a set threshold of events or thresholds in a system’s log.
Stakeholders	Distribution System Operator (DSO), grid operators, flex providers, aggregators

List of Figures

List of Tables

DIN e. V.

Am DIN-Platz
Burggrafenstraße 6
10787 Berlin
Tel.: +49 30 2601-0
Email: presse@din.de
Website: www.din.de

DKE German Commission for Electrical, Electronic & Information Technologies of DIN and VDE
Merianstraße 28
63069 Offenbach am Main
Tel.: +49 69 6308-0
Fax: +49 69 6308-9863
Email: dke@vde.com
Website: www.dke.de

December 2022