



# GERMAN STANDARDIZATION ROADMAP ON ARTIFICIAL INTELLIGENCE

Supported by:



Federal Ministry  
for Economic Affairs  
and Energy

on the basis of a decision  
by the German Bundestag

**PUBLISHED BY**  
**Wolfgang Wahlster**  
**Christoph Winterhalter**

**DIN**

**DIN e.V.**

Burggrafenstr. 6  
10787 Berlin  
Tel: +49 30 2601-0  
Email: [presse@din.de](mailto:presse@din.de)  
Website: [www.din.de](http://www.din.de)

**DKE**

**DKE German Commission for Electrical,  
Electronic & Information Technologies of DIN and VDE**

Stresemannallee 15  
60596 Frankfurt am Main  
Tel: +49 69 6308-0  
Fax: +49 69 08-9863  
Email: [standardisierung@vde.com](mailto:standardisierung@vde.com)  
Website: [www.dke.de](http://www.dke.de)

Photo credit:

Cover: LightFieldStudios – [istockphoto.com](https://www.istockphoto.com)  
Chapter entry pages: kras99 (p. 9, 33), assistant (p. 23),  
Thitichaya (p. 27), Maxim (p. 35), Shutter2U (p. 61),  
gunayaliyeva (p. 75), peshkov (p. 91), LuckyStep (p. 111),  
kaptn (p. 121), ryzhi (p. 129), Alex (p. 135),  
pickup (p. 143) – [stock.adobe.com](https://www.stock.adobe.com)

November 2020

## FOREWORD



Left: Prof. Dr. Dr. h.c. mult.  
Wolfgang Wahlster  
Head of the Steering Group,  
CEA DFKII

Right:  
Christoph Winterhalter  
Chairman of the Executive  
Board, DIN

Dear Reader,

With the Standardization Roadmap Artificial Intelligence, Germany is now the first country in the world to present a comprehensive analysis of the current state of and need for international standards and specifications for this key technology. In this first edition of the German Standardization Roadmap, not only the technical, but also the ethical and social aspects of standards in AI are taken into account in detail in a broad, interdisciplinary approach.

One of the twelve fields of action of the Federal Government's AI strategy of 2018 has been implemented with the preparation of this Roadmap; this strategy provides for a joint project with DIN for this purpose under Field of Action 10 "Setting Standards". DIN and DKE officially launched work on the Roadmap on behalf of the Federal Ministry of Economic Affairs and Energy (BMWi) at a kick-off event on 16 October 2019 with over 300 participants from industry, science, civil society and politics.

This is an ongoing document that needs to be regularly updated to reflect the enormously dynamic development of AI technologies and their rapidly expanding fields of application. Although all previously published standards and specifications in the field of AI are documented and the numerous ongoing standardization activities are shown in the Roadmap, many "white spots" on the AI standardization map have been identified which need to be filled in the next version.

The Roadmap was drawn up in seven working groups which developed important questions and recommendations for action on ethics, quality/conformity assessment/certification and IT security as horizontal topics, in addition to the basic

principles and the three AI application fields of particular importance for Germany – industrial automation, mobility/logistics and medicine.

AI is currently spearheading digitalization, because for the first time AI makes it possible to automate numerous cognitive services that could previously only be provided by human intelligence. AI systems are realized as pure software or as cyber-physical systems, but they always have to be linked to other current IT components in order to be applicable in practice. We therefore consider AI in the context of other digitalization trends such as cloud, edge, GPU and quantum computing, the Internet of Things and 5G, Industrie 4.0 and the platform economy.

Since the first wave of digitalization, most data have been machine-readable, as this resulted in the comprehensive replacement of analogue information processing and the almost complete digital capture, storage, transmission and storage of data. Numerous standards and specifications have helped to achieve this. But the second wave of digitalization, triggered as a driver by a wide range of AI technologies, is leading into the new era of machine-understandable data. Here digital data is interpreted, classified, enriched with meta-data and refined by AI systems in order to be able to draw new conclusions, develop new types of proposals for decisions, or achieve a goal set by humans through autonomous behaviour. However, we are still at the beginning of the necessary standards and specifications for this new era of digitalization, which our Roadmap documents for the first time, combining this with recommendations for the next steps.

Human-centred AI was at the forefront of the development of the Roadmap, which calls for all AI systems not only to be explanatory, robust and resilient, but also to take strict account of European values such as freedom from discrimination and the protection of privacy. Overall, standardization makes a decisive contribution to technical sovereignty and interoperability for AI applications, which will be of great relevance to all industries in the future.

It will be expedient to use AI technologies for standardization itself in the future, i.e. to apply document analysis, knowledge representation and machine learning to the creation, distribution and use of standards in order to move from machine-readable standards to machine-interpretable and verifiable standards.

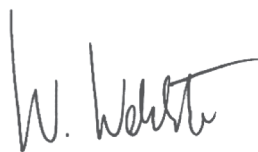
This approach offers the potential for a further significant increase in the annual savings of 17 billion euros already achieved today through standards in Germany. Standards accelerate the transfer of results of excellent AI research to the German economy and open up international markets, especially for medium-sized and start-up companies.

The preparation of this first Standardization Roadmap AI would not have been possible without the tireless efforts of our volunteer experts. The steering group consisting of twenty high-ranking personalities held six meetings in 2020 and adopted the final recommendations for action by consensus in a closed meeting, together with the heads of the seven working groups.

On behalf of the steering group, we would also like to take this opportunity to thank all 183 authors and 89 other participants for their great commitment. We would like to give our special praise to Ms. Filiz Elmas as the excellent coordinator of the overall project.

It is now important to implement as many of the recommendations for action as possible with the support of all federal ministries responsible for the AI strategy – which is currently funded by the amount of € 5 billion – and to prepare the ground today for the continuation of our Standardization Roadmap AI. The Roadmap shows that there is still a considerable need for research, e.g. to establish the necessary quality metrics and test profiles for risk-adapted certification of AI components.

We wish all readers an exciting read and ask for your active support in the further development of this Standardization Roadmap. Let us work together to develop and introduce international standards and specifications that support the safe use of “AI made in Germany” according to European values.



Prof. Dr. Dr. h.c. mult. Wolfgang Wahlster  
Head of the Steering Group, CEA DFKI



Christoph Winterhalter  
Chairman of the Executive Board, DIN

## GREETING



Peter Altmaier  
German Federal Minister for Economic Affairs and Energy

Dear Reader,

The use of voice assistant systems and image recognition show us that artificial intelligence is already reality today. And the list of further fields of application is long: It ranges from autonomous driving and intelligent traffic systems to medical diagnostics and therapy – all the way to industrial automation. With our approach of developing and using AI technologies in a responsible, human-centric manner with a focus on the common good, we strive to make Germany a leading AI location and to set the tone at the European level. In order to position ourselves to compete even more effectively on the international stage for the best ideas in the future, we have to set the right course today in these areas:

- regulatory, in which lawmakers create a regulatory framework for the development and use of AI technologies that promotes innovation;
- societal, by conducting a dialogue on the opportunities, risks and ethical issues associated with the use of AI technologies and
- technical, by describing uniform requirements with standards and specifications that support the implementation of the legal framework and ethical values.

Standardization has an essential role to play in achieving these objectives. Standards and specifications ensure interoperability, increase user-friendliness, and form the basis for trust in technical systems and processes. At the same time, they make it easier for German industry, dominated by small and medium-sized enterprises (“Mittelstand”), to access international markets and thus increase competitiveness. Two years ago, in November 2018, the Federal Government adopted its “National Strategy on Artificial Intelligence”

and identified “setting standards” as one of the twelve central fields of action. This newly released “German Standardization Roadmap on Artificial Intelligence” is a first step towards implementing the measures comprised in this field of action. The Roadmap describes the environment in which AI standardization operates, identifies existing standards and specifications relevant to the field of AI, and outlines further standardization needs. In addition, it formulates concrete recommendations for action which are aimed primarily at standardization actors, but also at stakeholders in quality infrastructure and policy.

The Standardization Roadmap makes it clear that we are all called upon to make “AI – Made in Germany” a model of success. It is not the conclusion, but rather just the beginning of the implementation of the measures from the field of action “Setting standards”. Industry, civil society, science and the public sector are therefore now challenged to work together to implement the recommendations of the Standardization Roadmap and to actively participate in shaping the rules of the game for the digital economy and society of the future. I am convinced that artificial intelligence requires European values. Let’s take on this challenge together!

A handwritten signature in black ink, appearing to read 'Peter Altmaier', written in a cursive style.

Peter Altmaier  
German Federal Minister for Economic Affairs and Energy

## Summary

DIN and DKE spent about a year working on the German Standardization Roadmap Artificial Intelligence in a joint project with the Federal Ministry for Economic Affairs and Energy and together with some 300 experts from industry, science, the public sector and civil society. A high-level steering group chaired by Prof. Wolfgang Wahlster coordinated and accompanied this work.

The aim of this Roadmap is the early development of a framework for action for standardization that will support the international competitiveness of German industry and will raise European values to international level.

The Standardization Roadmap AI implements a key measure of the German government's AI strategy, in which one of twelve fields of action is explicitly dedicated to the topic of "setting standards".

Standards and specifications play a special role particularly in the field of artificial intelligence: They promote the rapid transfer of technologies from research to application and open international markets for companies and their innovations. By defining requirements for products, services or processes, they ensure interoperability and quality. Standards and specifications thus contribute significantly to explainability and security and support the acceptance and trust in AI applications.

The present Standardization Roadmap AI was developed in a broad participation process with interdisciplinary actors, and outlines the work and discussion results of the working groups. It provides a comprehensive overview of the status quo, requirements and challenges for the following seven main topics:

- Basic topics
- Ethics/Responsible AI
- Quality, conformity assessment and certification
- IT security (and safety) in AI systems
- Industrial automation
- Mobility and logistics
- AI in medicine

The current environment of AI standardization for these central topics is described and an overview of relevant standards and specifications on aspects of artificial intelligence is given.

With over 70 identified standardization needs, the Roadmap shows concrete potential and formulates five central and overarching recommendations for action:

### 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.

### 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 procedures 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.

### 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.

#### 4. **Initiate and implement the national implementation programme “Trusted AI” to strengthen the European quality infrastructure**

So far, there is a lack of reliable quality criteria and test procedures 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 procedures 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.

#### 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 applications 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 requirements can be identified at an early stage and marketable AI standards realized.

The results of the Standardization Roadmap AI represent the prelude to the upcoming work and thus pave the way for future standardization in the field of artificial intelligence. Its implementation will help support German industry and science and create innovation-friendly conditions for the technology of the future. In particular, the results will make an important contribution to the socio-political debate at European level on the future role and use of AI.

Only early and comprehensive involvement of German stakeholders in national, but above all European and international standardization will strengthen Germany's position as an industrial nation and export country and pave the way for “AI – Made in Germany”.

The Standardization Roadmap AI will be continuously updated and developed to take account of changing requirements.

The task now is to launch concrete standardization activities along the lines of the recommendations for action. Interested experts are expressly invited to participate and contribute their knowledge in standardization.

<b>Foreword</b>	.....	<b>1</b>
<b>Greeting</b>	.....	<b>3</b>
<b>Summary</b>	.....	<b>4</b>
<b>1</b>	<b>Introduction</b> .....	<b>9</b>
<b>1.1</b>	<b>Trends in artificial intelligence</b> .....	<b>11</b>
<b>1.2</b>	<b>Standards for AI: four practical examples</b> .....	<b>13</b>
<b>1.3</b>	<b>Role of standardization in the field of AI</b> .....	<b>15</b>
<b>1.4</b>	<b>AI strategy of the German Federal Government</b> .....	<b>15</b>
<b>1.5</b>	<b>Objectives and content of the Standardization Roadmap AI</b> .....	<b>17</b>
<b>1.6</b>	<b>High-level steering group</b> .....	<b>18</b>
<b>1.7</b>	<b>Methodical approach</b> .....	<b>21</b>
<b>2</b>	<b>Recommendations for action of the Standardization Roadmap AI</b> .....	<b>23</b>
<b>3</b>	<b>Actors and the standardization environment</b> .....	<b>27</b>
<b>3.1</b>	<b>Socio-political environment</b> .....	<b>28</b>
<b>3.2</b>	<b>Innovative political initiatives</b> .....	<b>28</b>
<b>3.3</b>	<b>Standardization environment</b> .....	<b>29</b>
<b>4</b>	<b>Key topics</b> .....	<b>33</b>
<b>4.1</b>	<b>Basic topics</b> .....	<b>35</b>
4.1.1	Status quo .....	38
4.1.2	Requirements, challenges .....	38
4.1.3	Standardization needs .....	59
<b>4.2</b>	<b>Ethics/Responsible AI</b> .....	<b>61</b>
4.2.1	Status quo .....	62
4.2.2	Requirements, challenges .....	62
4.2.3	Standardization needs .....	73
<b>4.3</b>	<b>Quality, conformity assessment and certification</b> .....	<b>75</b>
4.3.1	Status quo .....	77
4.3.2	Requirements, challenges .....	81
4.3.3	Standardization needs .....	89
<b>4.4</b>	<b>IT Security (and safety) in AI systems</b> .....	<b>91</b>
4.4.1	Status quo .....	93
4.4.2	Requirements, challenges .....	96
4.4.3	Standardization needs .....	107
<b>4.5</b>	<b>Industrial automation</b> .....	<b>111</b>
4.5.1	Status quo .....	113
4.5.2	Requirements, challenges .....	114
4.5.3	Standardization needs .....	117
<b>4.6</b>	<b>Mobility and logistics</b> .....	<b>121</b>
4.6.1	Status quo .....	122
4.6.2	Requirements, challenges .....	123
4.6.3	Standardization needs .....	126



<b>4.7</b>	<b>AI in medicine</b> .....	<b>129</b>
4.7.1	Status quo .....	130
4.7.2	Requirements, challenges .....	130
4.7.3	Standardization needs .....	132
<b>5</b>	<b>Requirements for the development and use of standards and specifications</b> .....	<b>135</b>
<b>5.1</b>	<b>Review and development of standards and specifications in AI</b> .....	<b>136</b>
5.1.1	Review of existing standards and specifications .....	136
5.1.2	Agile development of standards and specifications for AI .....	136
<b>5.2</b>	<b>SMART standards – New design of standards for AI application processes</b> .....	<b>136</b>
5.2.1	Motivation .....	136
5.2.2	Status quo .....	137
5.2.3	SMART standards – Level model .....	139
5.2.4	Standards and AI .....	140
5.2.5	New design of standards for AI application processes .....	141
5.2.6	Summary and Outlook .....	142
<b>6</b>	<b>Overview of relevant documents, activities and committees on AI</b> .....	<b>143</b>
<b>6.1</b>	<b>Published standards and specifications on AI</b> .....	<b>144</b>
<b>6.2</b>	<b>Published standards and specifications with relevance to AI</b> .....	<b>147</b>
<b>6.3</b>	<b>Current standardization activities on AI</b> .....	<b>155</b>
<b>6.4</b>	<b>Committees on AI</b> .....	<b>161</b>
<b>7</b>	<b>List of abbreviations</b> .....	<b>165</b>
<b>8</b>	<b>Sources and bibliography</b> .....	<b>169</b>
<b>9</b>	<b>List of authors</b> .....	<b>189</b>
<b>10</b>	<b>Further working group members</b> .....	<b>195</b>
<b>11</b>	<b>Annex</b> .....	<b>199</b>
<b>11.1</b>	<b>Glossary</b> .....	<b>200</b>
<b>11.2</b>	<b>Philosophical foundations of ethics</b> .....	<b>204</b>
<b>11.3</b>	<b>SafeTRANS Roadmap</b> .....	<b>205</b>
<b>11.4</b>	<b>SMART Standards – New design of standards for AI application processes</b> .....	<b>206</b>
11.4.1	Use of granular content by means of the technology approach .....	207
11.4.2	Bottom-up method – Post-processing of standards .....	212
11.4.3	Top-down method – Development of SMART standards .....	213





**1**

# Introduction

Artificial intelligence (AI) has been ubiquitous for several years now, and today's digital world would be unthinkable without it. It is increasingly permeating more areas of social and economic life and will change the way we work, learn, communicate and consume.

Today there are numerous applications and existing practical examples of AI.

AI-based systems play a very present role in everyday life, e.g. when online retailers advertise additional products when users shop on the Internet, streaming services recommend new music playlists or films, social media platforms point out news, or smart watches detect cardiac arrhythmia or guide drivers to free parking spaces in real time.

The importance of AI is also growing rapidly in industrial applications. Experts assume that in the future AI will have such a great influence on industrial value creation that companies will hardly be able to refuse to use AI. The possibilities are almost limitless: including language assistants and chat bots, programs for document research, systems for diagnostic image recognition of tumors, industrial robots interacting with people in the factory, or autonomously driving cars.

AI is already widely used in companies today to optimize processes and increase productivity. These are mainly analytical activities that support decision-making processes. The great advantage of AI: It learns to produce better results than processes that follow rigid patterns and enables productivity and sales gains through increasingly personalized offerings. It therefore represents a technology that can be used to drive progress and secure the economic strength of Germany and thus the prosperity of an entire society.

The EU expects its economy to grow by 14 percent within the next ten years with the help of AI [1]. According to estimates, AI could increase the gross domestic product in Germany by 11,3 percent by 2030, which corresponds to a value added of 430 billion euros [2]. Not least for this reason, the European Commission and the German government have declared this technology a top priority (see [Chapter 1.4](#)).

The rapid increase in available data is seen as the main reason for the rise of this technology. Both the amount of data produced and the available computing power are increasing exponentially. AI lives on data; the more data are processed, the greater the potential learning effect and the more diverse the social benefits [1]. Among the top 10 most valuable

companies in the world are seven companies that make their money mainly with data. Only one German company whose business model is based on data appears in the top 100 [3].

Germany enjoys an excellent reputation in AI research. Many German research institutions and networks<sup>1</sup> belong to the global top AI research and have a clear knowledge advantage, for example in industrial AI applications in production areas. However, when it comes to developing innovative products and services from the research results and ultimately leading them to commercial success, other countries such as China or the USA are much more successful.

One thing is certain: If it is possible to combine the wealth of industrial experience of the German economy with the possibilities of data-driven AI methods to create an industrial AI, Germany could become a winner of the new AI technology, and could secure – and even expand – its competitiveness in the industrial sectors it already dominates.

However, some German companies have very different current situations, especially when it comes to the use of AI solutions: Some only know AI as a catchword, others have recognized the potential of AI technologies, but do not know where to start. Still others are planning the introduction of AI solutions, but are struggling to implement them. Over 99 percent of all companies in Germany are small and medium-sized enterprises who generate over half of the total value added. Thus, German SMEs in particular should see AI as a key technology and use its potential for themselves [4].

If “AI – Made in Germany“ is to be established as a brand and export hit in the future, AI technology must be considered an integral part of our economy today.

However, a technology will only be used successfully across the board if it finds acceptance in society. While companies are increasingly recognizing the opportunities offered by AI, the public discussion on this topic in Germany is very controversial. AI is sometimes rejected by the population due to ethical concerns.

Although AI methods are per se neither more neutral nor more discriminatory than human beings, they can still pro-

1 Examples are the European research associations ELLIS and CLAIRE, with which Germany has further consolidated its strong international position in research and development.

duce problematic or discriminatory decisions. AI systems are trained with data and information that are usually collected and processed by humans. If social prejudices or distortions are contained in these data, the AI system takes on these prejudices or even reinforces them in some cases, since an AI system has no moral judgement [4].

A clear framework for action is therefore needed to ensure that ethical values are respected. This is exactly where standards and specifications can be applied and help to increase the broad acceptance of AI systems by defining quality metrics, for example, thus making the reliability of the results of AI systems more assessable.

Beyond that, there is still much to be done, especially with regard to the security, fairness, robustness, transparency and adequacy of AI systems and their decisions. What is missing is a defined scope of action in which AI systems act for people and are based on transparent decision paths. The European Commission has set up a High-Level Expert Group on Artificial Intelligence (AI HLEG) for this purpose (see 3.1), which, among other things, has drawn up “Ethics guidelines for trustworthy AI” [5] as guidance for companies. If Germany succeeds in integrating European value standards into AI applications, German AI products will be able to gain greater acceptance worldwide than comparable products from the USA or China, for example. This is how the pioneering role of the German economy can succeed.

The participation of all stakeholders with the involvement of interdisciplinary actors – e.g. from computer science, engineering, philosophy, psychology, sociology, law, politics, civil society and consumers – provides a solid basis for a human-centred orientation and development of AI systems.

This AI standardization roadmap sets a significant milestone for such interdisciplinary cooperation by establishing an open, transparent and sustainable exchange. Resulting bodies, platforms, fora or activities can act as a catalyst in technology development.

## 1.1 Trends in artificial intelligence

In recent years, various technological developments have given an enormous boost to artificial intelligence and opened the race for global technological leadership. In the meantime, the use of AI has established itself as a global trend that no economy and hardly any company can escape. The facets of

the emerging trends are manifold. Especially with the progress of AI technologies and increasing success in technology development, new fields of application and possibilities are added almost daily. Some of these trends are taken up in the following and presented as examples.

There have always been efforts towards human interpretations, reactions and behaviour through AI systems, which have become increasingly important in recent years. With **neuromorphic computing** a milestone has been reached. Already today, humanoid robots interacting with humans can increasingly respond to traditionally soft factors, such as human emotions. AI systems can also mimic human cognitive processes and incorporate them into work processes, enabling AI systems in various applications not only to prepare human decisions, but sometimes to make decisions themselves. As a result, not only simple, power-consuming or routine activities are made possible by machines, but also cognitively demanding and creative activities that were previously reserved for humans.

The complexity of calculations, decisions, interpretations etc. by an AI system increases with the possibilities, which requires ever larger data volumes and higher processing speeds. One answer to this is provided by technologies such as **quantum computers**, which should accelerate progress in the development of artificial intelligence. Quantum computers can not only improve the performance of information processing, but may also make the use of AI methods possible under certain circumstances. One example is **quantum machine learning (QML)**.

Parallel to IT hardware, the performance of AI systems is also constantly evolving in terms of IT software. An example are AI systems that deal with non-hierarchical data and knowledge structures, as well as uncertainty, and can also bring unstructured data into a response structure. In addition to statistical methods, **ontologies** are a central element in order to extract meanings from data, recognize environmental conditions, and automatically derive recommendations for action or actions of the AI system.

New technologies and new types of processing, decision-making and action processes (e.g. in neuromorphic computing) increasingly offer further possibilities. Thus, even without the widespread application of the technology, it can already be seen that the boundaries of today’s automated systems will open up to highly **autonomous systems** if ethical values are observed. Unlike automated systems, autonomous

systems select their means to a certain degree independently and make decisions to achieve a given goal based on recognition of the situation in which they find themselves. Low-level autonomous systems are already being used in industry to weigh up utility aspects (flexibility, resources, time, quality, sustainability) against costs, safety aspects and industry-specific aspects of changes in the world of work (recruitment, qualification, redundancy, etc.). In terms of end customers, development is already more advanced, as demonstrated by the smart home and service robotics technologies.

Another trend that is emerging and which will be the main characteristic of successful AI systems is **self-explanatory capability**. This is realized through explanatory components that can explain the results of the AI system and the processing steps on which they are based in an argumentative dialogue that is understandable for the respective user, context-dependent and at different levels of detail. The dynamically generated explanations are mostly verbal, but occasionally also graphical or multimodal. The first AI systems (including the Mycin system), which could explain their own inference processes to a user asking a “why” question, were implemented as early as the mid 1970s. Especially for medical diagnostic systems, this is a prerequisite for the acceptance of physicians who are responsible to the patient for the use of a diagnostic or therapeutic suggestion. With knowledge-based systems, such introspection on the symbolic level is much easier to realize due to the explicit models of the application domain than with model-free systems based

on statistical or neural machine learning methods. But even for AI systems based on neuronal deep learning there are now first approaches for elementary explanatory components (see [6]).

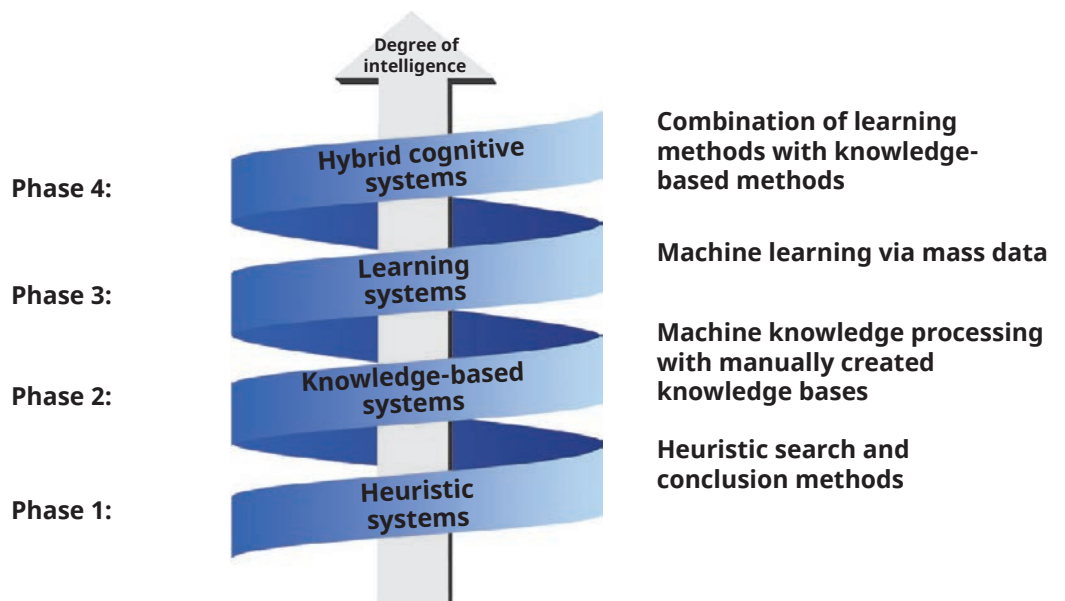
A global trend in AI is the development of **hybrid cognitive systems** that combine knowledge-based methods with machine learning, so that symbolic and sub-symbolic processes complement and reinforce each other. The global association for AI “AAAI” has clearly identified this trend for the USA in its roadmap for the next 20 years (see [7]). From today’s perspective, four phases of AI research (see Figure 1) can be distinguished, whereby hybrid cognitive systems currently exhibit the highest degree of intelligence, robustness, transparency and adaptability.

In the development of technology trends, a broad participation of stakeholders and interdisciplinary areas is becoming increasingly apparent. In addition to the involvement of new actors, there is sometimes a direct exchange with the end user, who thus has a stronger voice, and the development of AI systems takes a human-centred approach.

**Sustainability potentials through artificial intelligence**

In addition to the technological trends already mentioned, AI applications open up enormous possibilities, especially with regard to various aspects of sustainability. In the following, some areas of application are shown as examples.

**Figure 1:** The four phases of AI [8]



- In agriculture, in combination with drone- or sensor-based monitoring, for example, AI can help to assess the condition of plants and consequently to use fertilizers and pesticides in a more targeted and economical way (“precision farming”).
- In production, energy consumption can be reduced through networking and robotics.
- In the use phase, product life can be extended by means of predictive maintenance.
- In recycling and waste management, AI can improve the identification and sorting of waste, thereby increasing process efficiency and promoting recycling management.
- For building efficiency and energy management, AI offers the possibility of improved system control, both with regard to the regulation of heating, cooling and ventilation systems and the handling of networked production machines, especially when including IoT activities [9].

However, when it comes to the question of how sustainable AI and its applications really are, not only the field of application must be considered, but also the energy required for calculations. Since some computing power is very energy-intensive, it must be ensured that the most energy-efficient variant of the analysis is chosen.

All in all, therefore, AI can contribute considerably to greater sustainability if it is used correctly and ecological, economic and social aspects are taken into account.

## 1.2 Standards for AI: four practical examples

### Example 1: Standardized quality comparison of AI systems for automatic translation

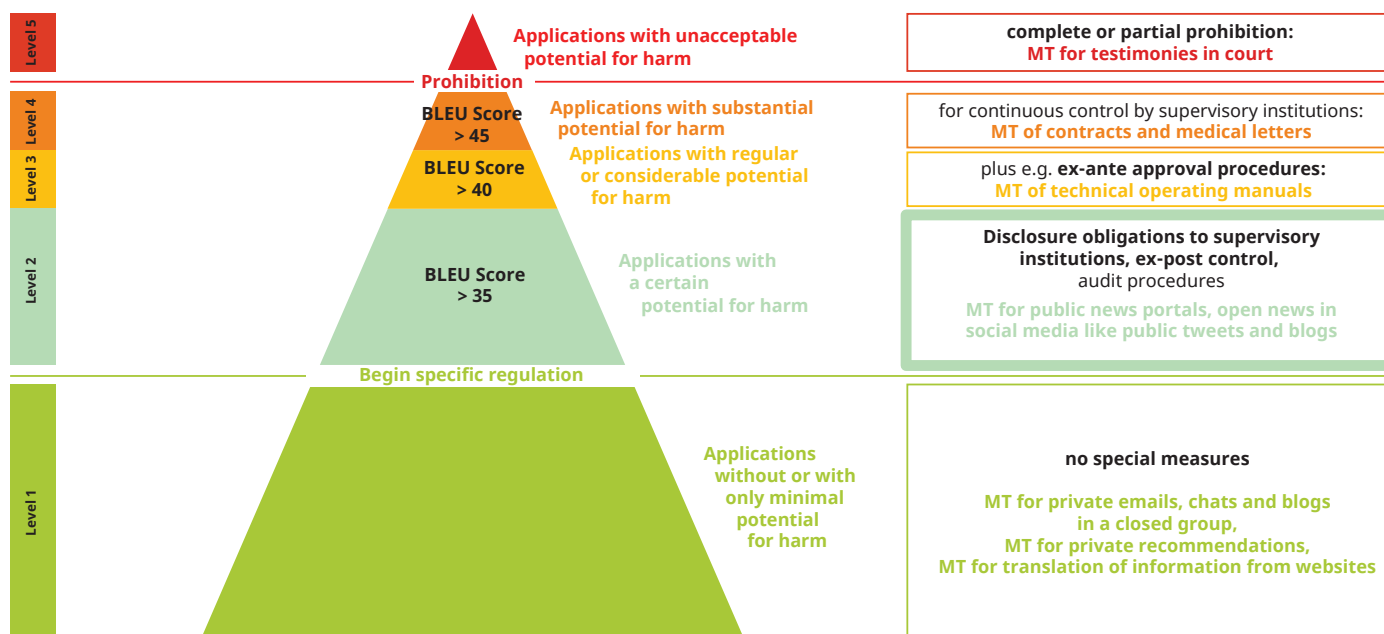
One of the oldest AI fields of application is the machine translation (MT) of texts. Already in 1954 IBM, together with Georgetown University, had programmed a minimal experimental system for Russian-English with only 250 words and six syntax rules. In the 70s, important basic modules for linguistically sound automatic translators for German were developed in Germany in the Collaborative Research Centre 100 in Saarbrücken with the MT system SUSY. Today MT systems are widely used. They are offered by Internet companies such as Google, Microsoft, Facebook, Amazon, and Baidu, but also by a very successful German AI start-up called DeepL, and have billions of users every day in daily life and at work. Currently, MT systems based on neural translation algorithms are the most widespread. These are trained over millions of

pairs of sentences in the source and target languages often simply over the character strings end-to-end, in so-called transformer architectures with a sequence of encoders for input and another sequence of decoders for output. Previously, statistical methods were the most successful, while purely symbolic methods, despite their greater precision in special domains (e.g. weather reports), receded into the background due to their insufficient coverage for an everyday language.

In the field of machine translation, the gold standard for measuring the accuracy of a translation is the BLEU value (bilingual evaluation understudy), which automatically compares machine translated texts with variants of texts translated by human experts. Although the BLEU value (on a scale of up to 100 %) does not capture syntactic and semantic correctness due to its very simplified metric, which is limited to word sequence comparisons, it has proven to be useful for a rough quality assessment and performance comparison between MT systems. The standardization of quality metrics for AI systems is a very important issue for the broad acceptance of such systems in practical use, since the reliability of the results can be better assessed. NIST has developed a slightly modified BLEU factor that has been used since 2006 in the annual tests of MT systems for different translation domains in the annual machine translation workshops and conferences (WMT) to evaluate system performance. For example, a BLEU value of 15 is a very bad value that requires a lot of effort in the post-editing of the automatic translation to be able to continue working with the translated text.

Peak values, which are currently achieved by the best MT systems, are currently just over 45. However, the primitive BLEU metric cannot be used to evaluate the severity of translation errors, e.g. if a negation at the wrong place distorts the entire statement in the source text.

For a risk-adaptive certification of AI systems, quality barriers must be defined for certain application classes, below which the results of the AI system cannot be further used in a critical application context (see Figure 2). If, for example, a witness statement is available in a foreign language, its machine translation by an AI system with a too high error rate of course cannot be used in court, but must be produced by a sworn human translator (red area in the criticality pyramid). Only certified MT systems should be used for operating manuals for technical equipment, and the quality of MT for contract texts and doctor’s reports must be continuously checked. If there is a systematic suspicion of manipulation through content-falsifying translations in public tweets, blogs or news



**Figure 2:** The criticality pyramid and a risk-adapted regulatory system for the use of AI-based machine translation systems (according to [10])

portals, an ex-post control of the MT system used should be possible. On the other hand, an AI translation service with a lower BLEU value may be useful in the context of translating private chats and recommendations, because there is a very low risk for the user if the translation is incorrect.

**Example 2: Semantic AI technologies to ensure the interoperability of systems**

In Industrie 4.0, production machines from different manufacturers communicate with each other. Machine tool builders often use different terminology. Machine-understandable ontologies from AI make it possible for different systems of concepts to be automatically transferred into each other, so that a communication between different machines in the Internet of Things (IoT) becomes possible. This requires standardized ontology description languages. With a W3C consortium standard called OWL (Ontology Web Language), a standard was created with the significant participation of German AI experts, which today is also used in German industry. Many companies have already specified their own terminology systems for their production machines with the help of OWL.

**Example 3: AI for action planning**

AI-based action planning is an area in which Germany is regarded as very successful in international research. In simple

terms, it is a matter of finding a sequence of individual action steps starting from an initial state that leads to a desired target state. Such AI planning systems are, for example, an important AI technology for planning the actions of autonomous robots, transport planning in logistics, or production planning in the Smart Factory. Here a de facto standard called PDDL (Planning Domain Definition Language) and, recently, for hierarchical planning the variant HDDL (Hierarchical Domain Definition Language) have established themselves as the specification languages for, among other things, the preconditions and postconditions of an action step which is used, for example, for the global competition and comparison of the best planning systems, IPC 2020. German AI planning systems have already won several times in these global competitions, which have been held since 1998. In Industrie 4.0 and in autonomous driving, AI planning systems are among the components that are critical to success.

**Example 4: Standardization in the area of data-driven machine learning (ML)**

With AI methods of deep learning, very good progress could be made in the field of automatic analysis of images and image sequences on the basis of multi-layer neural networks. However, learning success depends not only on the quantity of training data, but also on its quality. Especially in supervised learning, extensive training data often first has to be



annotated by human experts, e.g. which dog breed is shown on one of 100,000 images in the training data set. In order to obtain data sets that are as valid as possible, several annotators are commissioned to process overlapping data sets. Here, their consistency in the assessment must be checked. The **Kappa statistics** were introduced as a gold standard for this purpose. Thus, as one of the criteria for the quality of the annotated training data, the reliability of the annotations for a certification of an AI system based on ML becomes operationalizable and comparable through a standard metric.

### 1.3 Role of standardization in the field of AI

The ability to implement new ideas and research findings as products and services is decisive for the competitive ability of German industry. Standardization can serve as a catalyst for innovations, and helps bring solutions to the market sustainably.

Standards and specifications define requirements for products, services or processes and thus lay the foundation for technical procurement and product development. At the same time, standards and specifications ensure interoperability and serve to protect people, the environment and property, and to improve quality in all walks of life. In this way, they create transparency and trust in the application of technologies, and at the same time support communication between all parties involved by using uniform terms and concepts. Standards generate economic benefits which have been estimated at € 17 billion a year for Germany alone [11].

Standards and specifications play a very special role in creating a sustainable framework for artificial intelligence: They promote the rapid transfer of technologies from research to application and open international markets for German companies and their innovations. AI is an area where the full (and timely) commitment of the German stakeholders in standardization at national level, and above all at European and international level, can play a decisive part in reinforcing Germany's position as a leading economy and export nation.

German SMEs in particular can benefit from this. This is a major advantage of standardization. The following principle applies: it is not the larger party that decides, but the consensus. Participation gives innovative small and medium-sized companies the opportunity to work on the future of AI on an equal footing with the large national and international corpo-

rations and to contribute their own ideas to the standardization process. Open interfaces and uniform requirements give them better access to the global market and the opportunity to position their ideas there.

Such a commitment is extremely important from a national and European perspective: Those who enforce their standards internationally have a head start because their own rules apply and can be built on existing solutions. Germany's competitors are aware of this, especially China and the USA. These nations naturally pursue their very own interests – and their ideas may conflict with our European values and ethical guidelines. However, the fact that the question of technical sovereignty and, above all, data sovereignty is being pursued, especially in value-oriented Germany and Europe, is demonstrated by lighthouse projects like GAIA-X, which are intended to manifest the added value of “AI – Made in Germany” in an international context. Standards support sovereignty by promoting transparency and setting framework conditions that provide a “moral compass”. Although it is the task of society and politics to define what is ethical, technical standards can help to implement existing ethical values and thus ensure protection in a technical context, for example against distortion, discrimination and manipulation.

In this context, standards make a significant contribution to explainability and traceability – two essential building blocks when it comes to the acceptance of AI applications. At the same time, standards ensure security and engender trust, aspects of crucial significance in a field as sensitive as AI. The German government also assigns a central role to standards, especially in the field of artificial intelligence. Not least for this reason, standardization is a central component of the German Federal Government's AI strategy.

### 1.4 AI strategy of the German Federal Government

On 15 November 2018 the Federal Government adopted the national strategy “Artificial Intelligence” [12], thus accelerating the path of “Artificial Intelligence – Made in Germany” to being a world leader. With this strategy, the Federal Government aims to secure Germany's excellent position as a location for research, to expand the competitiveness of German industry and of Europe, and to promote the diverse applications of AI in all areas of society. The benefits for humans and the environment is the focus of attention, and the intensive exchange on the topic of AI with all social groups

will be strengthened. In order to achieve these ambitious goals, the German government has decided, as part of its latest economic stimulus package [13], to increase the planned investments for AI promotion from three billion euros to five billion euros by 2025. The focus is on research, transfer, social dialogue, qualification and data availability. This will support a competitive European AI network.

The main objectives of the Federal Government’s AI strategy are:

- Strengthening the competitive ability of Germany and Europe
- The responsible development and use of AI for the common good
- Embedding AI into society in an ethical, legal, cultural and institutional context

The strategy describes concrete measures in 12 fields of action (see Figure 3).

Standardization is one of the twelve fields of action. In Field of action 10 “Set standards” the goal is:

“In a joint project with DIN, the German government will (among other things) develop a roadmap on standards and specifications in the field of AI.”

In addition, the review of existing standards and specifications for «AI suitability» and the development of machine-readable and machine-interpretable standards and specifications (smart standards) for AI applications is suggested (see Chapter 5).

**AI strategies of other countries**

The race for the world’s leading position in artificial intelligence has long since begun. Since then, many countries and economic areas have sought ways to promote research and the application of AI in their countries/regions and have developed their own national strategies for this purpose. Below, AI strategies of individual countries and their special features are presented as examples [14]:

**European Commission White Paper on AI**

With its “White Paper on Artificial Intelligence: a European approach to excellence and trust” [15], the EU Commission has published its vision for a safe and responsible use of artificial intelligence. It represents a first attempt to establish clear rules on what AI may and may not do, and suggests approaches on how to enforce them. The focus is on making AI usable for science, industry and society, while at the same time addressing the associated risks. The proposed measures include, for example, increased cooperation with and between Member States, and the pooling of expertise by facilitating the establishment of centres of excellence and testing.

**Figure 3:** The twelve fields of action of the Federal Government’s AI strategy



## USA

The USA's leading position in AI can be summed up by the following figures: it is the country with the highest number of AI publications worldwide (around 22,000), has around 2,400 AI start-ups and thus is the world's largest AI start-up landscape, is number one in the use of AI applications in companies (25 % of companies), and is home to seven of the world's ten largest technology companies, and also home to cooperation structures between universities, public authorities and companies that have grown over the last 40 years [4]. Given these factors, it is not surprising that the Obama administration also presented the world's first national AI strategy as early as 2016.

In addition, DARPA (the Defense Advanced Research Project Agency of the USA) has recently launched a new \$2 billion funding initiative, called "AI Next" for a period of five years to develop the foundations for the next generation of AI systems [16]. DARPA aims to promote a new wave of AI systems that are more robust and trustworthy than previous systems because they are based on a tight integration of components for perceiving, learning, context understanding, inferring and planning, and on an explicit representation of the knowledge used in problem solving. DARPA's aim is to overcome AI development, which the US government believes has, in many countries, recently focused too much on machine learning and to develop a new generation of autonomous systems that can also work in teams with humans. In a "third wave of AI", AI systems are thus to be transformed from pure tools into real collaborative partners for concrete problem solving.

## China:

In three steps, China plans to become the leading AI nation in the world by 2030 and is setting economic targets for this move. Over 700 million Chinese Internet users and powerful hardware and technology groups are good prerequisites for this. Although the country still lags behind the USA in basic research, the training of qualified specialists, the number of AI start-ups and international patents, developments in recent years leave no doubt that China is catching up. Beijing has announced 16,4 billion euros to promote the chip industry alone, and at the subnational level a single city (Tijian) has set up a fund of 12,8 billion euros for AI promotion. With the "Thousand Talents" program, Beijing also wants to attract highly qualified foreign Chinese back to the nation. However, despite the massive deployment of funds, a scientific breakthrough cannot be planned, especially in the face of weak basic research. In addition to capital, this requires above all a conducive academic environment.

## United Kingdom

In early 2018, the British government and the private sector agreed to jointly fund research and commercialization of AI with one billion euros. Apart from the internationally very influential AI research, the main strength of the country is the AI start-up scene. Nowhere else in Europe are more AI start-ups concentrated. At the same time, the government has laid the foundations for the development of ethical guidelines for AI, for example by founding a Centre for Ethics. In recent years, Great Britain has expanded its technology cooperation with the USA. On the other hand, the country has weaknesses in the commercialization of research, which is manifested, among other things, by the low number of patents.

## France

France is formulating a claim to leadership based on a middle ground between China and the USA that is founded on European values. In AI development, the country relies on a centralized structure and organization of the state system. The responsible ministries focus their strategies and resources on AI applications in the fields of health, mobility and defence. Structural weaknesses can be seen in the low number of institutes and teaching staff actively researching in areas directly related to AI (Great Britain has almost eight times more, Germany about four times more) and the lack of cooperation between universities and industry. In planned AI centres of excellence, scientists will therefore be pooled on the one hand in order to be able to work with users with a certain degree of autonomy. On the other hand, France is setting new rules that allow researchers to work in the private sector at the same time. A network of voluntary AI experts will advise the state in the procurement of technologies and support cybersecurity.

## 1.5 Objectives and content of the Standardization Roadmap AI

The early development of a framework for action which sets out requirements in the field of standardization is essential and necessary. The resulting impetus for corresponding work in standardization at national, but above all at European and international level, can decisively strengthen Germany's role as an economic nation and export country.

Especially in such a sensitive subject area as AI, decisive steps can be taken that lead to trust and security in the use of AI. The subject of AI inevitably also has a recognizable link to legislation. In this process, the requirements laid down

in standards and specifications can at the same time have a relieving effect on legislation and thus also contribute to an acceleration of the establishment of framework conditions, for example.

In order to achieve a leading role in this process, it is important to position ourselves accordingly. The basis for this must be a coordinated approach to the relevant subject areas and a functioning network. The development of corresponding recommendations in the form of a standardization roadmap can make a significant contribution towards introducing the national position on the basis of a broad coordination process at European and ultimately at international level. DIN and DKE provide a recognized and neutral platform for orchestrating this work with the help of their many years of expertise and network competence.

In order to develop a framework for standardization in the field of AI at an early stage, DIN and DKE initiated work on the standardization roadmap “Artificial Intelligence” on behalf of the Federal Ministry of Economics and Energy (BMWi). With this step, the AI strategy of the Federal Republic is implemented (see 1.4). On 16 October 2019, a kick-off event attended by more than 300 participants from industry, civil society, politics and science gave the starting signal for the AI standardization roadmap.

The standardization roadmap is a “living document” which presents the results of work and discussions to date and serves as a central communication medium for exchange between standardization bodies, industry, associations, research institutions and politics.

The aim of the AI standardization roadmap is to describe at an early stage a framework for action that will strengthen German industry and science in the international competition for the best solutions and products in the field of artificial intelligence, and create innovation-friendly conditions for the technology of the future.

It identifies needs for standards and specifications, especially with regard to the security, reliability and robustness of AI systems, and contributes significantly to ensuring the quality of AI solutions. In addition to describing the environment in which the actors operate, it provides an overview of existing standards and specifications on aspects of AI, outlines the main potential for standardization, and makes recommendations for action to policy-makers, researchers and standardizers. The Roadmap thus makes a significant contribution to

establishing “AI – Made in Germany” as a strong brand and developing new business models, trailblazing innovations and scalable applications. At the same time it offers great potential for raising European values to the international level.

The AI standardization roadmap will be developed and regularly updated in an open, transparent and broad-based participation process by representatives from industry, science, public authorities and society.

Based on the results of the standardization roadmap and the identified recommendations for action (see [Chapter 2](#)), concrete standardization activities will be communicated to the relevant standards committees as a next step.

### 1.6 High-level steering group

The AI Standardization Roadmap is being overseen by a group of high-ranking representatives from industry, politics, science and civil society. Prof. Wolfgang Wahlster, Member of the Steering Committee for the Platform Learning Systems (PLS) and leading German AI research scientist, is heading up the steering group.

The 20-member group is responsible for the content and strategic orientation of the Standardization Roadmap AI, paving the way for the expansion of Germany as an AI location. The members represent important topics, disciplines, industries and companies of different sizes in the field of AI and see themselves as ambassadors for the transfer of scientific results through standards to the economy and important areas of life.

With the founding of the steering group, an important step has been taken in creating the necessary framework for artificial intelligence.



Photos: Impressions from the kick-off event

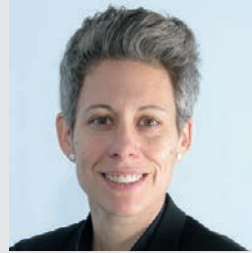
MEMBERS OF THE  
STEERING GROUP



**Dr. Tarek R. Besold**  
neurocat GmbH



**Jörg Bienert**  
Federal Association  
Artificial Intelligence



**Dr. Julia Borggräfe**  
Federal Ministry of  
Labour and Social Affairs



**Dr. Joachim Bühler**  
TÜV Association



**Susanne Dehmel**  
Bitkom e. V.



**Dr. Dirk Hecker**  
Fraunhofer – Alliance  
Big Data and Artificial  
Intelligence



**Thorsten Herrmann**  
Microsoft Deutschland  
GmbH



**Stefan Heumann**  
Stiftung Neue  
Verantwortung



**Dr. Wolfgang Hildesheim**  
IBM Germany



**Prof. Jana Koehler**  
German Research Center  
for Artificial Intelligence



**Prof. Klaus Mainzer**  
Technical University of  
Munich



**Dr. Christoph Peylo**  
Robert Bosch GmbH



**Alexander Rabe**  
eco Association of the  
Internet Industry



**Prof. Ina Schieferdecker**  
Federal Ministry of  
Education and Research



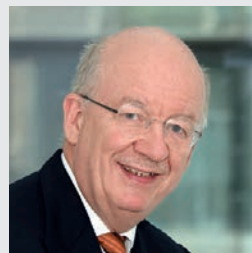
**Stefan Schnorr**  
Federal Ministry for  
Economic Affairs and  
Energy



**Andreas Steier, MdB**  
Christian Democratic  
Union Germany



**Dr. Volker Treier**  
German Chamber of  
Commerce and Industry



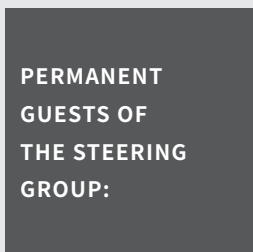
**Prof. Wolfgang Wahlster**  
Steering Group Platform  
Learning Systems



**Prof. Dieter Wegener**  
Siemens AG



**Christoph Winterhalter**  
DIN German Institute for  
Standardization



PERMANENT  
GUESTS OF  
THE STEERING  
GROUP:



**Dr. Gerhard Schabhüser**  
BSI Federal Office for  
Information Security



**Dr. Johannes Winter**  
Platform Learning  
Systems

## 1.7 Methodical approach

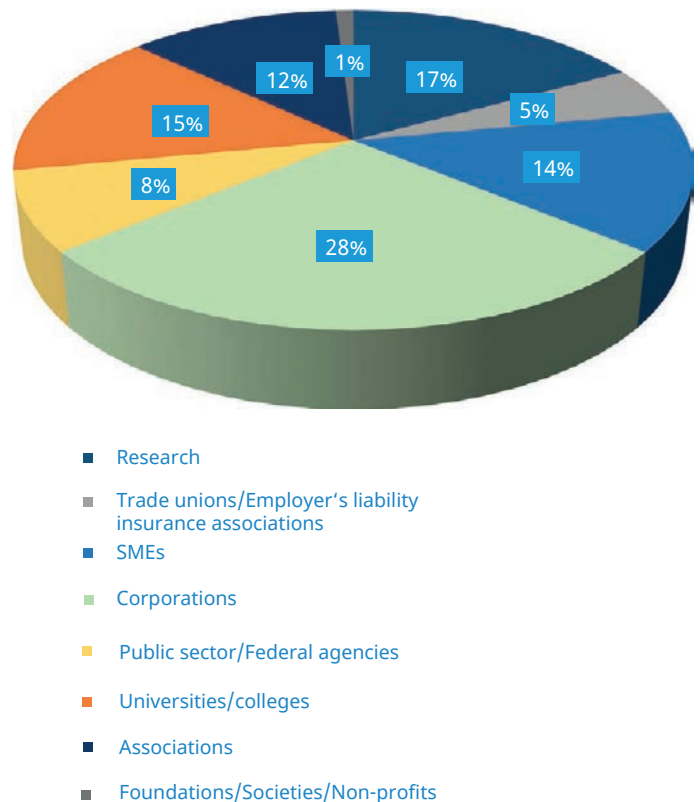
The participation of experts from all relevant areas is the essential basis for drawing up the standardization roadmap. Stakeholders involved include industry representatives from the relevant sectors, experts from the scientific community, representatives from politics and civil society, as well as representatives of already constituted groups concerned with the topic of AI. The consideration of different points of view and associated requirements is of great importance here, so that both technical and non-technical aspects have been equally incorporated into the development process of the Standardization Roadmap AI.

The development of the Standardization Roadmap AI involved the overall coordination and orchestration of the relevant stakeholders and took place in seven working groups on various key topics (see [Chapter 4](#)). Experienced experts were recruited to lead the working groups, who led the content work and reported to the steering group:

1. Basic topics (Head: Dr. Peter Deussen, Microsoft Germany GmbH, and Dr. Wolfgang Hildesheim, IBM Germany)
2. Ethics/Responsible AI (Head: Tobias Krafft, Technical University Kaiserslautern)
3. Quality, conformity assessment and certification (Head: Dr. Maximilian Poretschkin, Fraunhofer Institute for Intelligent Analysis and Information Systems IAIS, and Daniel Loevenich, Federal Office for Information Security)
4. IT security in AI systems (Head: Annegrit Seyerlein-Klug, secunet Security Networks AG)
5. Industrial automation (Head: Dr.-Ing. Christoph Legat, HEKUMA GmbH)
6. Mobility and logistics (Head: Dr. Reinhard Stolle and Bogdan Bereczki, both Argo AI)
7. AI in medicine (Head: Prof. Dr. Johann Wilhelm Weidringer, Bavarian State Chamber of Physicians and German Medical Association)

Around 300 experts from various sectors and with different backgrounds contributed their expertise to the seven working groups. [Figure 4](#) shows the composition of the working groups.

The content was developed on the digital working platform [DIN.ONE \(www.din.one/site/ki\)](http://www.din.one/site/ki).



**Figure 4:** Composition of the seven working groups of the Standardization Roadmap AI

The Standardization Roadmap AI will be presented at the Digital Summit 2020 and handed over to the German government. It can be downloaded for free in German and English at [www.din.de/go/normungsroadmapki](http://www.din.de/go/normungsroadmapki).

The publication of the Standardization Roadmap AI will be immediately followed by its implementation. This means that on the basis of the results of the Roadmap and its recommendations for action (see [Chapter 2](#)), concrete standardization activities will be initiated as the next step.







## 2

# Recommendations for action of the Standardization Roadmap AI

The aim of the Standardization Roadmap AI is to describe at an early stage a framework for action that will strengthen German industry and science in the international competition for the best solutions and products in the field of artificial intelligence and create innovation-friendly conditions for the technology of the future. The Roadmap thus makes a significant contribution to establishing “AI – Made in Germany” as a strong brand and developing new business models, groundbreaking innovations and scalable applications. In particular, German SMEs and the growing start-up scene in Germany can benefit if it is possible to combine the wealth of industrial experience of the German economy with the possibilities opened up by AI methods. 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 – essential building blocks when it comes to the acceptance of AI applications. The Standardization Roadmap AI thus offers great potential for both securing Germany’s competitiveness and raising European value standards to the international level. Not least for this reason should special attention be paid to the implementation of the present Standardization Roadmap AI and its recommendations for action.

#### **Recommendation for action 1: Data reference models for the interoperability of AI systems**

A large part of German industry is made up of small and medium-sized enterprises. The realization of overarching value chains therefore often requires the cooperation of the most diverse actors. The automation of the collaboration of AI systems of different actors along the value chain is crucial for the application of artificial intelligence methods. This requires a data reference model to enable secure, reliable, flexible and compatible data exchange between technologies. In such a model, data types that are basic and relevant for interoperability, and their structures and relationships to each other should be described.<sup>2</sup> The Standardization Roadmap AI therefore recommends an implementation program for the standardization of data reference models in different domains. Through such an initiative, Germany can create the basis for a comprehensive exchange of data by playing a key

role in shaping international standards, thus ensuring the interoperability of systems worldwide.<sup>3</sup>

#### **Recommendation for action 2: Development of a horizontal AI basic security standard**

In essence, an AI system is an IT system for whose IT security a multitude of standards from various industries and fields of application already exist. At the same time, such a variety increases complexity and a lack of transparency, which can lead to an inconsistent approach of the market players involved (e.g. manufacturers, consumers, regulators) and significantly inhibit the technological development of AI. IT security for AI systems in particular suffers from this, as it depends to a large extent on transparency and traceability, security-by-design, security-by-default and privacy over the entire life cycle. An all-encompassing “AI umbrella standard”, which bundles existing standards and test methods for IT security (security, safety and privacy) and supplements them with aspects specifically for AI systems, can on the one hand serve as a catalyst for technology development, and on the other hand mediate between the actors. For this reason, the creation of a horizontal basic security standard for AI is recommended, which considers further topics and industries with their specifics in vertical sub-standards and integrates an IT (security) management system for AI systems. This would maintain established procedures and create a testable and certifiable standard that takes economic aspects into account and increases acceptance. This in turn builds trust and promotes the use of AI technologies.<sup>4</sup>

#### **Recommendation for action 3: Practice-oriented initial criticality checking of AI systems**

Unintended ethical problems and conflicts occur primarily in ADM systems with learning components that make decisions about people, their belongings or access to scarce resources, and have the potential to damage individual fundamental rights and/or basic democratic values. An initial criticality check as to whether a system can trigger such conflicts at all or whether it is an application far removed from any ethical issue, must be made quick and easy by standardization. This horizontal, for all areas low-threshold check must quickly and legally clarify whether the system must meet transparency and traceability requirements at all. Especially with regard to the wide fields of application of artificial intelligence, such a

<sup>2</sup> See Example 2: Semantic AI technologies to ensure the interoperability of systems

<sup>3</sup> Principles and aims of such an implementation program are explained in 4.3.2.4.

<sup>4</sup> For more information see 4.4.2.3.

risk-based criticality check in critical areas offers the possibility to make adequate demands and at the same time to counter the accusation of “ethical red taping” by developing completely uncritical fields of application free of additional requirements. Therefore, the design of a practical and risk-based criticality check for AI systems is recommended.<sup>5</sup>

#### **Recommendation for action 4: National implementation program “Trusted AI” to strengthen the European quality infrastructure**

Industry, public authorities and civil society demand reliable quality criteria and test procedures for the marketable conformity assessment and certification of AI systems. The lack of such test procedures endangers the economic growth and competitiveness of this future technology. At the same time, statements about the trustworthiness of AI systems cannot be substantiated without high-quality test methods, which means that the benefits of AI applications to the economy and society remain unclear due to lack of acceptance. A successful use that meets our ethical and social values requires competent, reliable and reproducible tests. The basis for this can be a national implementation program for the certification of AI systems, which builds on the excellent German testing infrastructure and defines requirements, for example, for reliability, robustness, performance and functional safety. On the basis of concrete application cases, testing principles are to be tested, pilot tests carried out and standards derived which form the basis for AI certification and are to be introduced into international standardization. The test methods to be developed serve on the one hand to confirm the assured properties of AI systems (product testing) and on the other hand to evaluate the measures taken by organizations providing AI systems (management system testing). With such an initiative, Germany would have the chance to lay the foundation for the world’s first certification program and thus become a leader in the development and standardization of an internationally recognized AI certification procedure. The Standardization Roadmap AI therefore recommends the fastest possible initiation and implementation of a national implementation program “Trusted AI” with the highest priority.<sup>6</sup>

#### **Recommendation for action 5: Analyze and evaluate use cases for standardization needs**

Use Cases describe application cases that are essential for understanding the function and behaviour of AI systems

in the context of AI technologies. There are already a large number of use cases for different fields of application of AI. By considering application-typical and sector-relevant use cases, standardization needs can be derived for industrially mature AI applications. However, the proven approach of traditional standardization is not always appropriate for AI applications. The reason is that many industries use different AI technologies depending on the field of application of the AI solution and relating to the use case. Hybrid AI solutions are often even based on a combination of AI methods. In most cases, the specifics of the application are met by state-of-the-art approaches from AI sub-disciplines, which are individually adapted and refined. Consequently, the dynamics at the interface between AI research and industrial development and application are particularly high. Thus, the applied AI is constantly being developed and industrially evaluated. AI standardization must take this tension between applied research and industrially mature development into account and pursue pragmatic, bidirectional approaches in the analysis of standardization needs and the development of market-ready specifications. This requires an iterative process which, in the design of standards and specifications, incorporates reciprocal impulses from research, industry, society and regulation and supports continuous and mutual learning between the actors. At the centre of this approach is the testing and successive refinement of the developed specifications along use cases. In this way, application-specific requirements can be identified at an early stage and marketable AI specifications can be realized. As a result, the acceptance of AI specifications by industry, science and society is ensured.<sup>7</sup>

5 For more information see [4.1.2.1.5](#).

6 For more information see [4.3](#).

7 For more information see [4.5.2.2](#).



The background is a complex, abstract digital landscape. It features a central circular motif composed of concentric rings and overlapping rectangular frames, resembling a stylized 'A' or a data visualization. This central element is surrounded by a dense network of white lines and nodes, some of which are highlighted with small white circles. The overall color palette is monochromatic, using shades of gray and white against a dark background.

**3**

## Actors and the standardization environment

There are currently a large number of actors, initiatives, committees and standardization activities dealing with the topic of AI at national, European and international level. In the following, the AI environment with the main actors and initiatives is presented<sup>8</sup>.

### 3.1 Socio-political environment

#### Ethics Commission on Automated and Connected Driving

The Federal Ministry of Transport and Digital Infrastructure's "Ethics commission on Automated and Connected Driving" was set up in September 2016. The interdisciplinary commission included high-ranking experts from philosophy, law and social sciences, technology assessment, consumer protection, the automotive industry and the digital economy. It was the first committee in the world to address the important socially relevant issues in automated and connected vehicular traffic. In its final report, the ethics committee has drawn up a total of twenty ethical rules or "development guidelines" [17].

#### Enquete Commission

The Enquete Commission (commission of enquiry) "Artificial Intelligence – Social Responsibility and Economic, Social and Ecological Potentials" [18] was appointed by Germany's parliament, the Bundestag, in June 2018 to investigate the future influence of AI on social life, the economy and the world of work – all areas in which standardization also has a major influence. The Commission is made up equally of members of the German Bundestag (in percentage representation of the respective parliamentary group in parliament) and external experts. The members work in six project groups, in each case three of which meet or met in parallel:

- AI and industry (industry/production, finance, services, innovations)
- AI and the State (administration, security, infrastructure)
- AI and health (care, sport)
- AI and work, education, research
- AI and mobility (energy, logistics, environment)
- AI and the media (social media, opinion-making, democracy)

For the project groups AI and industry, AI and the State, and AI and health, summaries of the preliminary results were already published in December 2019 [19]–[21]. The final report of the Enquete Commission is expected in autumn 2020.

#### High-Level Expert Group on Artificial Intelligence

The High-Level Expert Group on Artificial Intelligence (AI HLEG) is composed of 52 experts from science, civil society and industry and is the central body of the European Commission in the field of AI. Its task is to support the implementation of the European AI strategy. This includes the development of recommendations for future policy development and ethical, legal and societal issues related to AI, including socio-economic challenges.

The AI HLEG has presented the following results in 2018 and 2019:

- Ethics guidelines for trustworthy AI [5]:  
The guidelines put forward a human-centric approach to AI and list seven key requirements that AI systems should meet in order to be trustworthy. They cover topics such as fairness, security, transparency, future of work, democracy, privacy and protection of personal data.
- Policy and investment recommendations:  
Building on its initial findings, the AI HLEG made 33 recommendations to strengthen Europe's competitiveness, including guidelines for a strategic research agenda on AI and for the establishment of a network of AI centres of excellence [22]. The recommendations will help the Commission and Member States to update their joint coordinated plan on AI. This is expected to play a key role in building the future of artificial intelligence in Europe.

### 3.2 Innovative political initiatives

#### Plattform Lernende Systeme (Platform Learning Systems)

The Plattform Lernende Systeme (Platform Learning Systems) (PLS) was initiated in 2017 by the German Federal Ministry of Education and Research with the aim of shaping AI for the benefit of society. It brings together some 200 AI experts from science, industry, politics and civil society. In seven working groups (WGs) they develop options for action and recommendations for the responsible use of learning systems, five of which show parallels to the topics of this standardization roadmap:

- PLS WG 1 "Technologies and Data Science"
- PLS WG 2 "Work and Skilling and Human-Machine Interaction"

<sup>8</sup> The presentation makes no claim to completeness.

- PLS WG 3 “IT Security, Privacy, Law and Ethics”
- PLS WG 4 “Business models”
- PLS WG 5 “Mobility”
- PLS WG 6 “Medicine and Care”
- PLS WG 7 “Hostile-to-Life Environments”

In its publications, the PLS analyses technological, economic, moral and social conditions for the responsible and self-determined use of AI systems in various application areas (e.g. medicine and mobility). It also examines cross-cutting issues such as discrimination, certification or the IT security of AI systems. Using industry-specific application scenarios, the PLS shows what will be technologically possible with AI in a few years’ time and what general conditions need to be created. On its “Map on AI”, it shows where AI is already being used in Germany and which institutions are conducting research on the topic. The combination of all activities of the PLS represents an intersection with standardization, which results in standardization potentials.

#### Platform Industrie 4.0

The Platform Industrie 4.0 (PI4.0) represents the central network for advancing the digital transformation in industrial value creation. Founded in 2013 by the trade associations BITKOM, VDMA and ZVEI, it now comprises over 350 players from companies, associations, trade unions, science and politics. Relevant aspects of Industrie 4.0 are being considered in currently six working groups. A research advisory board brings scientific, research and development expertise to the working groups and gives impulses regarding future research topics.

In PI4.0, AI is considered a cross-sectional topic. As a result, an “Artificial Intelligence” project group was founded within the platform with the aim of considering this topic in terms of its application and thematic embedding across working groups and providing corresponding impulses in the existing working groups [23], [15]. The project group completed its dedicated work in the first quarter of 2020. Further work will now be continued in the six working groups.

- PI4.0 WG 1 “Reference Architectures, Standards and Standardization” [25]
- PI4.0 WG 2 “Technology and Application Scenarios”
- PI4.0 WG 3 “Security of Networked Systems” [26], [27]
- PI4.0 WG 4 “Legal Framework” [28]
- PI4.0 WG 5 “Work, Education and Training” [29]
- PI4.0 WG 6 “Digital Business Models in Industrie 4.0”

#### Platform Future of Mobility

As an initiative of the Federal Ministry of Transport and Digital Infrastructure (BMVI), the [National Platform Future of Mobility](#) (NPM) supports the German government in achieving its goals, for example in the transport sector and climate protection. In detail the NPM has the following overarching goals:

- Develop multi-modal and intermodal solutions for a largely greenhouse gas-neutral and environmentally friendly transport system
- Ensure a competitive automotive industry and promote Germany as an employment location
- Enable efficient, high-quality, flexible, safe, resilient and affordable mobility for persons and goods

AI systems are essential for the future and full implementation of these goals. Conventional IT systems have already reached their limits due to the complexity of optimization and processing, for example. In order for standardization to support the goals of the Federal Government and the NPM, WG 6 “Standardization, norms, certification and type approval” of the NPM is the link between the organizations. This enables a direct exchange between technical rule setters, legislators, industry and research.

### 3.3 Standardization environment

The essential standardization work is a joint task which is carried out in self-regulation by the stakeholders (such as industry, science, research, users, consumer protection, occupational health and safety, trade unions, public authorities and environmental protection). The starting point is always a need on the part of the stakeholders.

The development of standards and specifications takes place on a variety of levels (national, European and international) (see [Figure 5](#)).

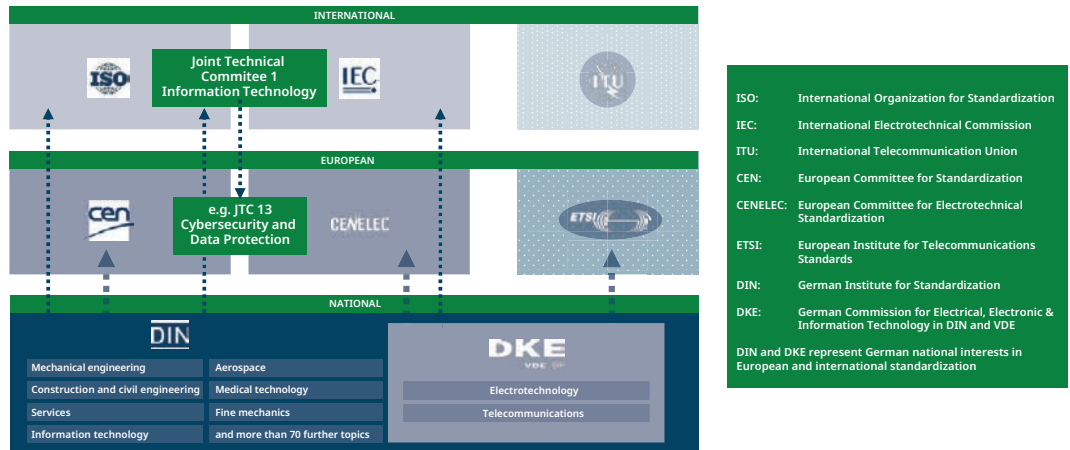
In Germany, DIN<sup>9</sup> has been contractually the responsible standards organization of the Federal Republic of Germany since 1975 and represents German interests as a member of CEN<sup>10</sup> in European standardization and of ISO<sup>11</sup> in interna-

9 German Institute for Standardization, [www.din.de](http://www.din.de).

10 Comité Européen de Normalisation, European Committee for Standardization, [www.cen.eu](http://www.cen.eu)

11 International Organization for Standardization, [www.iso.org](http://www.iso.org)

**Figure 5:** National, European and international levels of standardization



tional standardization. The DKE<sup>12</sup> represents German interests in electrical engineering, electronics and information technology in the field of international and regional electrical engineering standardization work. It thus represents German interests both at CENELEC<sup>13</sup> and in the IEC<sup>14</sup>.

Today almost 90 percent of DIN’s and DKE’s standards work is European and/or international in nature. DIN and DKE coordinate the entire standardization process at national level and ensure the participation of the relevant German national bodies at European and international level.

As technical rules, standards are the result of national, European or international standardization work and are developed by committees according to defined principles, procedures and rules of presentation<sup>15</sup>. All interested parties, such as manufacturers, consumers, the trades, universities, research institutes, authorities, testing institutes, etc., can participate in the work of the committees. Standards are developed by consensus. This means that experts come to agreement on the state of the art and on standards content that take the interests of all parties into consideration. According to this definition, all standardization documents of the national standards organizations (DIN/DKE), the European standards organizations (CEN/CENELEC/ETSI) and

the international standards organizations (ISO/IEC/ITU) are referred to as „standards“ in the context of this Roadmap.

In parallel, the general term „specifications“ refers to all other technical rules such as technical reports (TR), pre-standards, technical specifications (TS, DIN SPEC), consortium standards, application rules (AR), guidelines, expert recommendations, etc., for the preparation and publication of which the above-mentioned organizations as well as other organizations and technical rule setters may be responsible. For example, topics that have not yet fully arrived on the market or whose market does not yet exist are often dealt with in consortial standards or specifications. This may also be related to a low level of maturity (or „Technology Readiness Level“). In the case of specifications, consensus and the involvement of all stakeholders are not mandatory.

At present, work on standards and specifications on AI is being carried out at all levels.

At national level, standardization work on AI is being carried out in Germany within the **DIN Standards Committee Information Technology and selected IT Applications** (Working Committee NA 043-01-42 AA). This committee elaborates the German position in AI standardization and at the same time mirrors work at international and European level.<sup>16</sup>

12 DKE German Commission for Electrical, Electronic & Information Technologie in DIN and VDE

13 Comité Européen de Normalisation Électrotechnique, European Committee for Electrotechnical Standardization [www.cenelec.eu](http://www.cenelec.eu)

14 International Electrotechnical Commission, [www.iec.ch](http://www.iec.ch)

15 As a rule, the use of standards is voluntary. They only become mandatory if they are referred to in contracts, laws or regulations.

16 On behalf of the German Federal Ministry of Economics and Energy, DIN and DKE have developed a White Paper on “Ethical aspects in standardization for AI in autonomous machines and vehicles”, the results of which have been incorporated into the work of the present Roadmap.



At European level, the CEN/CENELEC Focus Group on Artificial Intelligence is a relevant body. It was established in 2019 by CEN and CENELEC as a temporary working group with the task of developing a roadmap for AI standardization at European level.

At international level, [ISO/IEC JTC 1/SC 42 “Artificial Intelligence”](#) is the central body for AI standardization and is therefore responsible for the development and publication of international standards on AI.

Apart from formal standardization, there are a number of professional associations and consortia that publish corresponding specifications or recommendations on AI. A considerable amount of the consortium work on AI standardization takes place within various forums and consortia such as IETF, IEEE, CSA, OGC, OMG and W3C.

**Chapter 6** gives a comprehensive overview of the main documents, activities and committees in standardization in the field of artificial intelligence.

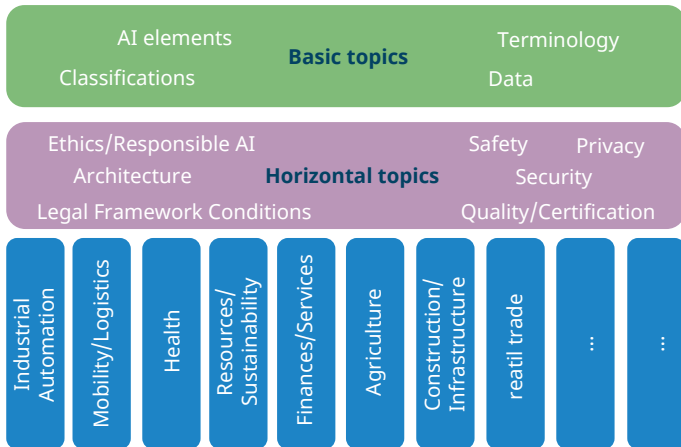




4

Key topics

Due to its scope and complexity, it seems reasonable to structure the topic of AI according to basic topics, horizontal topics, as well as relevant economic and application areas (see Figure 6).



**Figure 6:** Chart of basic topics, horizontal topics, and relevant economic and application areas

In this first version, the present Standardization Roadmap AI focuses on the areas of basic topics, horizontal topics (ethics, quality/conformity assessment/certification, IT security) and the three application areas of industrial automation, mobility/logistics and health.

In the following Chapter 4.1 to 4.7, the starting situation, challenges and essential standardization needs of the seven main topics are elaborated.

The basic topics form the basis for discussions on AI. This includes, for example, terminology (definitions), classifications, but also topics such as data (data analyses, data formats, data quality, etc.).

New technical developments, especially in the application of AI, raise new questions on overarching issues such as IT security, quality, ethics or the legal framework. Ethical aspects of responsibility in the use of AI technologies, as well as issues such as fairness, security, social inclusion and transparency of algorithms must be considered. In addition, the foundations for cross-industry quality criteria must be developed to enable the analysis and certification of AI systems. Which legal relationship AI may have in the future is another cross-sectional topic to be discussed.

The economic fields of application for AI are extremely diverse. AI is relevant for almost all sectors of the economy, and also for other areas of application outside the economy, and is found both in the form of components in end products and services, and in the productive core processes and support processes within companies.



## 4.1

## Basic topics

**Definition of the term “artificial intelligence”**

Providing a precise definition of the term “artificial intelligence” is a difficult task due to a multitude of different perspectives and opinions on this topic:

1. Does the term refer to a scientific or technical discipline, or does it refer to a system property or capability?
2. Should the term be limited to a description of the function of AI systems or refer to their implementation?
3. Should terms commonly associated with human intelligence (like “knowledge”, “skills”) be used to explain AI?

Almost every organization dealing with AI defines this term in different ways. In view of the difficulties in finding a generally accepted definition, this will not be done in this document.

4.1.2.1 gives an overview of the different classes of AI methods and their capabilities and areas of application, which will be used for the following discussion to narrow down the term. However, the range of possible definitions of AI is illustrated by the following examples in Table 1:

**Table 1:** Different definitions of AI

Example	German	English	Source
1	Künstliche Intelligenz beschreibt Systeme, die intelligentes Verhalten dadurch zeigen, dass sie – mit einem gewissen Grad an Autonomie – ihre Umgebung analysieren und entsprechend agieren, um spezifische Ziele zu erreichen.	Artificial intelligence (AI) refers to systems that display intelligent behaviour by analyzing their environment and taking actions – with some degree of autonomy – to achieve specific goals.	[30]
2	<System> Fähigkeit, sich Wissen anzueignen, zu verarbeiten, zu kreieren und anzuwenden, das in einem Modell gespeichert wird, um eine oder mehrere vorgegebene Aufgaben zu erfüllen <Technische Disziplin> Disziplin zur Entwicklung und Erforschung von KI Systemen <Künstliche Intelligenz> Informationen zu Objekten, Ereignissen, Konzepten oder Regeln, ihren Beziehungen und Eigenschaften, zur zielorientierten Nutzung organisiert Anmerkung 1 zum Begriff: Information kann in numerischer oder symbolischer Form existieren. Anmerkung 2 zum Begriff: Informationen sind kontextualisierte Daten, die damit interpretierbar werden. Daten werden durch Abstraktion oder durch Messungen der Umgebung kreiert.	<system> capability to acquire, process, create and apply knowledge, held in the form of a model, to conduct one or more given tasks <engineering discipline> discipline of developing and studying AI systems <artificial intelligence> information about objects, events, concepts or rules, their relationships and properties, organized for goal-oriented systematic use Note 1 to entry: Information may exist in numeric or symbolic form. Note 2 to entry: Information is data that has been contextualized, so that it is interpretable. Data are created through abstraction or measurement from the world.	ISO/CD 22989, ongoing project in ISO/IEC JTC 1/SC 42, currently at Committee Draft (CD) stage
3	Das Design und die Konstruktion intelligenter Agenten, die Wahrnehmungen ihrer Umgebung erhalten und deren Handlungen ihre Umgebung beeinflussen.	The designing and building of intelligent agents that receive percepts from the environment and take actions that affect that environment.	[31]
4	Ein KI-System ist ein maschinenbasiertes System, das in der Lage ist, für eine vorgegebene Menge von durch den Menschen definierte Ziele Vorhersagen, Empfehlungen oder Entscheidungen, die reale oder virtuelle Umgebungen beeinflussen, vorzunehmen. KI-Systeme werden entwickelt, um mit verschiedenen Graden von Autonomie zu operieren.	An AI system is a machine-based system that can, for a given set of human defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy.	[32]
5	Künstliche Intelligenz (KI) ist ein Teilgebiet der Informatik mit dem Ziel, intelligentes Verhalten und die zugrundeliegenden kognitiven Fähigkeiten auf digitalen Computern zu realisieren.	Artificial intelligence (AI) is a branch of computer science with the goal of realizing intelligent behaviour and the underlying cognitive abilities on digital computers.	[33]

Autonomous systems [33] can independently solve complex tasks in a specific application domain despite varying objectives and initial situations. Autonomous systems must independently generate an action plan, depending on the current task context, with which an overall goal specified by the operator of the autonomous system can be achieved without remote control and, if possible, without the intervention and assistance of human operators within the framework of legal and ethical requirements. If individual actions of the autonomous system fail during the execution of the plan, the system must be able to carry out a plan revision on its own in order to achieve the specified objective by adapting the original plan in another way. A new generation of autonomous systems is also able to solve a distributed task together with other autonomous systems and/or a group of people. Within the framework of self-regulation, an autonomous system must also have explicit models of its own performance limits and, in the case of specifications or environmental conditions that do not indicate a successful autonomous achievement of objectives, must inform the system operator of this fact (e.g. excessive shear winds prevent drone flight, an extremely steep section of the route exceeds the maximum climbing capacity of an autonomous vehicle).

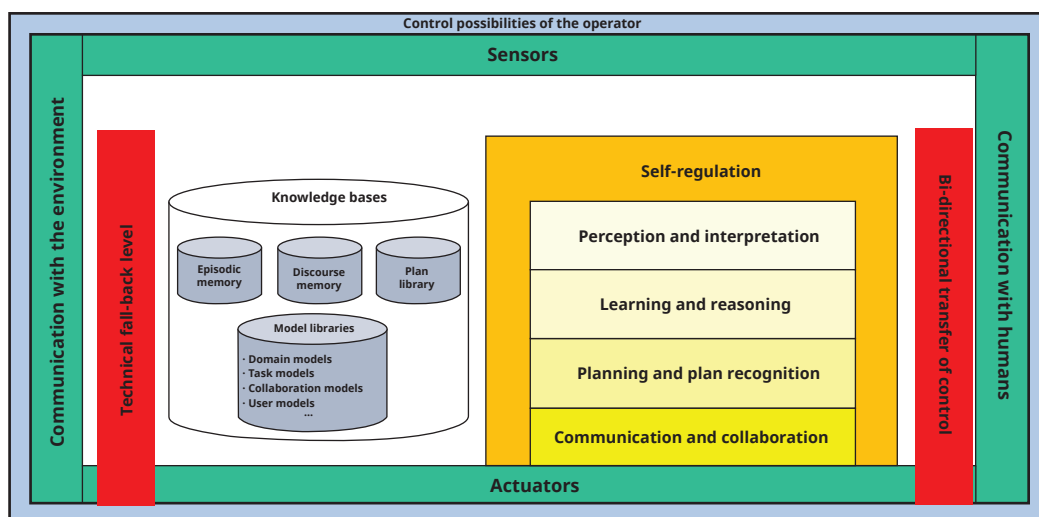
A reference architecture for autonomous systems has been developed in the High-Tech Forum of the German Federal Government (see Figure 7). It is based on sensors for observing the environment and actuators for changing environmental conditions in order to achieve the objectives of the autonomous system. In addition, communication with the networked environment of the system and with cooperating humans can provide further important information for

the behaviour of the autonomous system. In principle, the autonomous system consists of several modules for cognitive information processing, which are controlled by different mechanisms for self-regulation, as well as several knowledge bases, which are constantly adapted by machine learning and reasoning starting from an initial configuration.

With knowledge bases, an episodic memory serves as a long-term memory for events that have directly affected the autonomous system to enable case-based reasoning and learning from experience. The entire course of the system's communication with humans and technical systems in the environment is stored in the discourse memory, so that references to the aforementioned and ambiguities in the context can be resolved at any time. A plan library stores successfully executed plans for common classes of problems in order to achieve goals more efficiently through plan revision without replanning, and through plan recognition based on observing the actions of other agents in the environment to identify their intent.

Domain models contain networked models of all relevant objects, relations, states and events in an application field, which are necessary for their recognition by sensors or for their transformation by the actuators of the autonomous system. In task models, typical task classes for an autonomous system are schematically recorded in order to quickly understand and classify a new task set by the system operator or to decompose it into a series of known tasks. User models are particularly crucial when using autonomous systems as assistance systems in the service sector, since they contain assumptions about the preferences, abilities

**Figure 7:** Reference architecture for autonomous systems [33]



and level of knowledge of a system user, among other things, which enable personalization of service performance through adaptive behaviour.

In order to increase confidence in the use of autonomous systems and to minimize the risk of endangering people in the environment in the event of a complete technical failure of the central control functions, there must be a technical fall-back level in accordance with the reference architecture which, in an emergency, puts the autonomous system into a safe operating state, for example via a redundant mechatronic function or a radio-based remote control, and generates an alarm message via communication with the environment.

More often, an autonomous system will reach the limits of its abilities in abnormal situations and will have to hand over control to a human being. A bi-directional transfer of control must be provided for, so that after overcoming an obstacle to the achievement of objectives which cannot be achieved by the autonomous system alone, the human being can completely return control to the system.

#### 4.1.1 Status quo

With regard to AI basic topics, the SC 42 is carrying out work on various documents:

- **ISO/IEC 22989, Artificial intelligence – Concepts and terminology.** This standard is being developed under the leadership of a German editor.
- **ISO/IEC 23053, Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)** describes a terminological framework for machine learning.
- **ISO/IEC 23894, Information Technology – Artificial Intelligence – Risk Management** contains guidelines for the risk management for the development and use of AI systems. This standard, too, is being developed under the leadership of a German editor.
- **ISO/IEC 38507, Information technology – Governance of IT – Governance implications of the use of artificial intelligence by organizations** deals with organizational governance in connection with AI.
- **ISO/IEC 20546, Information technology – Big data – Overview and vocabulary [34]** deals with concepts and terminology relating to big data, which is also being considered within SC 42.
- **ISO/IEC 5059, Software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – Quality Model for AI-based systems**

Various Technical Reports give an overview of the current state of the art. These include:

- **ISO/IEC TR 24027, Information technology – Artificial Intelligence (AI) – Bias in AI systems and AI aided decision making**
- **ISO/IEC TR 24368, Information technology – Artificial intelligence – Overview of ethical and societal concerns.**

Projects on the following topics are currently being coordinated and are expected to start work in autumn 2020:

- A certifiable management standard for AI that contains requirements and organizations for the responsible development and use of AI systems.
- Various projects on the description of methods and processes for data quality in the context of machine learning. One of these projects is under the leadership of a German editor.

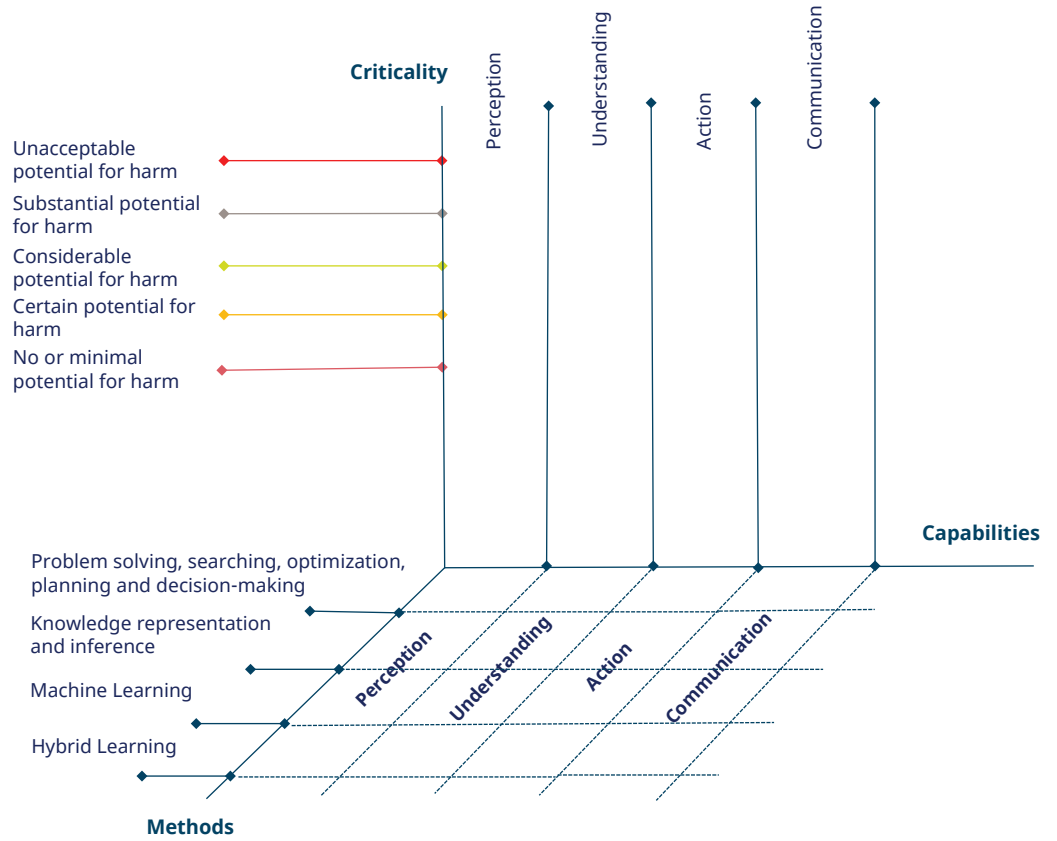
#### 4.1.2 Requirements, challenges

The evaluation of AI applications with regard to their suitability can be based on ethical, legal and technical criteria. In view of the progressively growing AI market, an overview of application scenarios as well as embedded methods (4.1.2.1) and capabilities (4.1.2.2) of AI is indispensable. This helps to avoid shortcomings in development, deployment, conformity assessment and the determination of quality characteristics of AI. While 4.1.2.1.3 gives an overview of applications with embedded methods and capabilities of AI within software markets, a classification of AI applications based on different degrees of decision autonomy is presented in 4.1.2.1.4. Besides a description of AI through methods, capabilities and degree of autonomy, aspects such as “right to privacy, “basic right to life and physical integrity” can be reflected through criticality (4.1.2.1.5) (see Figure 8).

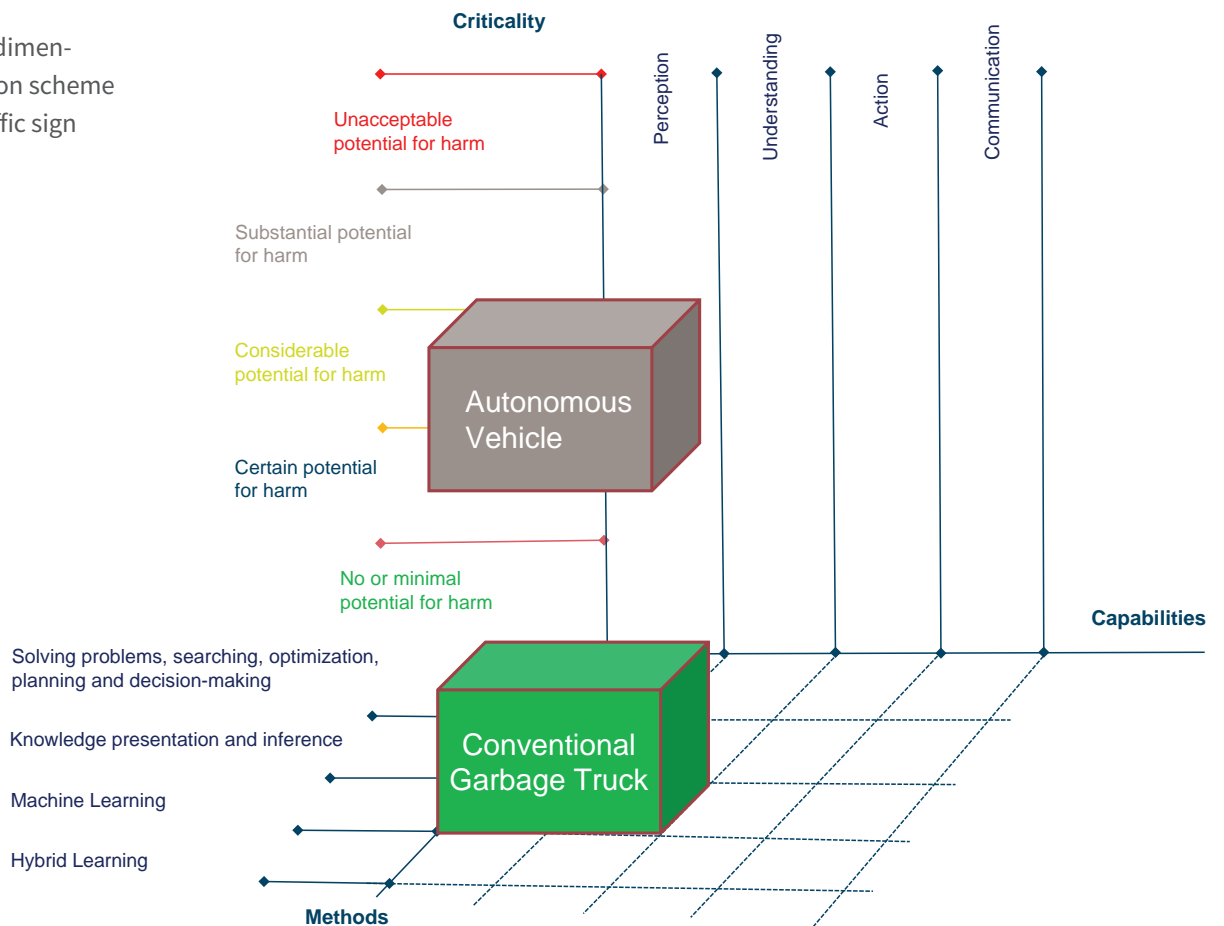
Given the wide range of capabilities of AI applications, the potential for harm plays a decisive role in the social acceptance of AI. Using the example of AI-based recognition of traffic signs, the potential for harm can vary depending on the application: In road traffic, a substantial potential for harm can be assumed for self-propelled motor vehicles due to the high amount of concerns and obligations. In contrast, a conventional garbage truck with the same AI technology for traffic sign recognition does not represent a self-propelled vehicle, so that a minimal potential for harm can be assumed (see Figure 9).



**Figure 8:** Three-dimensional classification scheme for evaluating an AI-based system



**Figure 9:** Three-dimensional classification scheme to evaluate AI traffic sign recognition



## Annotation of real images for training data for traffic sign recognition



Software tool for annotators: The German Traffic Sign Benchmark GTSRB

**Figure 10:** AI-based traffic sign recognition

A simple example of an AI application in a vehicle is a system for video-based traffic sign recognition. Here the detection of speed limits is already standard equipment in many cars. Since many traffic signs in connection with the permissible maximum speed have only a temporary validity (e.g. road works, gantries for dynamic traffic control), the necessary information cannot be taken from digital maps alone, but is recognized via pattern recognition using images from a camera, usually on the interior mirror. In this way, even recently erected signs, for example on construction sites, are registered. But that is not enough: Camera-based traffic sign recognition not only evaluates data based on signs. Instead, this recognition is compared with other assistance systems, such as the navigation system, the rain sensor and the time in order to correctly interpret restricted speed limits. However, driver assistance systems available on the market for traffic sign recognition do not work 100 % correctly, but a test showed a recognition rate between 32,5 % for the worst system and 95 % for the best system [35] on a course with 40 signs on speed limits which 12 cars passed. Temporarily invalidated speed signs using adhesive tape as well as speed displays in tunnels and illuminated signs on sign gantries proved to be a great challenge, as did the mix-up of speed limits for a turning lane (see Figure 10).

This simple example makes it clear that standards and test methods for this relatively simple subtask for autonomous driving according to Level 5 are necessary to ensure conformity of driving with road traffic regulations.

For this purpose, a standardized training data set for the traffic signs must be specified and benchmark tests must be provided for certification. In a risk-based approach, a detection rate of 99,9 % would have to be achieved for autonomous driving, while detection rates below 80 % also show considerable risks for product liability with pure assistance functions.

### 4.1.2.1 Classification

Following the position paper “A definition of AI: Main capabilities and scientific disciplines” of the AI HLEG [30], a distinction is made between methods and capabilities of AI. In both cases the following classifications are based on the standard work of Russell and Norvig [31] and integrate the current state of the art. The matrix in Table 2 shows which AI methods are used to realize certain AI capabilities. In order to also adequately reflect the actual state of the current industrial

AI markets, a classification of AI applications resulting from AI methods and AI capabilities is also carried out. Detailed information can be found in [Table 2](#) to [Table 5](#) and in the recently published Beuth Pocket [36].

#### 4.1.2.1.1 Classification of AI methods

The methods of AI generally move within a kind of spectrum between symbolic and sub-symbolic – sometimes also called numerical – methods. In terms of symbolic methods, there are especially techniques of knowledge representation and logical reasoning, while sub-symbolic methods are primarily represented by techniques of machine learning. In between are methods of problem solving/optimizing/planning/decision-making as well as hybrid learning methods that use both symbolic and sub-symbolic techniques.

Symbolic AI is especially characterized by a deductive procedure, i.e. by the (algorithmic) application of logical rules or relations to individual cases. A distinction is made between methods for representing knowledge on the one hand and methods for applying this knowledge on the other. Knowledge can be represented either as certain or uncertain. In knowledge application the classical methods of logical reasoning are suitable for certain knowledge. For reasoning based on uncertain knowledge, probabilistic approaches are widely used, but there are also a number of nonprobabilistic approaches.

Subsymbolic AI is characterized in particular by an inductive procedure, i.e. by the (algorithmic) derivation of general rules or relationships from individual cases. In most cases a distinction is made between supervised learning to achieve a given goal and unsupervised learning without a comparable goal. When both approaches are combined, it is referred to as partially supervised learning. In addition, there is also known reinforcing learning without fixed target parameters, which does not require a fixed target value, but qualitative specifications (right/wrong).

The method complex of problem solving/optimizing/planning/decision-making comprises algorithms and procedures that focus on these sub-areas. Examples are intelligent agents, methods of game theory and evolutionary algorithms.

Hybrid procedures are often characterized by the fact that they combine sub-symbolic with other AI techniques, e.g. to be able to work both inductively and deductively. In contrast to classical sub-symbolic procedures, a form of knowledge representation is often used additionally. In contrast to classical symbolic methods, however, such knowledge representations are often algorithmically modified depending on input data.

**Table 2:** Classification of AI methods

CLASSIFICATION OF METHODS ACCORDING TO TOPICS			EXAMPLES
PROBLEM SOLVING, SEARCHING, OPTIMIZATION, PLANNING, DECISION-MAKING	Problem solving	Problem solving agents, problem solving through searching, search strategies	Uninformed and informed search strategies Adversarial searching (game theory) Searching with boundary and secondary conditions (constraint solving)
	Optimization	Statistical optimization methods	Local search for optimization Searching in continuous spaces Searching with partial observation Searching in unknown environments Dynamic programming
		Bio-inspired optimization methods	Evolutionary algorithms Genetic algorithms / genetic programming Swarm intelligence
	Planning and plan recognition	Autonomous and semi-automatic planning methods	State space search Planning graphs Hierarchical planning Planning in non-deterministic domains Time and resource planning methods Generation of plans
		Plan recognition methods	Plan recognition via abductive reasoning Deductive plan recognition Recognition via plan libraries Recognition via plan synthesis
	Decision-making	Approaches to decision-making	Models Use / value of information Decision networks Decision-theoretical expert systems Sequential decision problems Iteration models

CLASSIFICATION OF METHODS ACCORDING TO TOPICS		EXAMPLES	
KNOWLEDGE REPRESENTATION AND INFERENCE	Representation of knowledge	Knowledge representation languages and models	RDF
			RDFS
			OWL
			KIF
		Structure and formality	
		Ontological engineering	Taxonomy
			Ontology
			Interpretation
			Calculus
			Deduction
			Abduction
			Ontology mapping
		Knowledge graphs and semantic networks	Knowledge networks / graphs
			Existence graph
			Graph traversing algorithms
			Mapping
			Semantic Web
		Modelling in formal logic	Propositional logic and predicate logic
			Higher-level logics, non-monotonic logics
			Temporal and modal logic
	Logical reasoning	Automatic proof methods	Resolution provers, connection provers
			SAT and SMT solvers
		Model checking	
		Interactive proof methods	Tactical theorem proving
	Uncertain knowledge	Quantifying uncertainty	Bayes's rule
		Representation of uncertain knowledge	Bayesian network
	Probabilistic reasoning	Inference in Bayesian networks	Exact inference
			Approximate inference
			Markov chain simulation
		Relational probability models	Relational probability models in closed/open universes
		Time and uncertainty in probabilistic reasoning	Hidden Markov model
		Kalman filter	
		Dynamic Bayesian networks	
	Non-probabilistic approaches	Qualitative approaches	Reasoning with default information
			Truth Maintenance Systems (TMS)
		Rule-based approaches	Rule-based reasoning with "certainty factor"
		Vagueness reasoning	Fuzzy quantities and fuzzy logic
		Reasoning with belief function	Dempster-Shafer theory
	Further approaches to uncertain reasoning		Spatial reasoning
			Case-based reasoning
			Qualitative physics
			Psychological reasoning

CLASSIFICATION OF METHODS ACCORDING TO TOPICS		EXAMPLES		
MACHINE LEARNING	Supervised learning	Neural networks	Multi-layer perceptron Learning Vector Quantization (LVQ) Radial basis function networks (RBF) Adaptive Resonance Theory (ART) Convolutional Neuronal Networks (CNN) Recurrent Neural Networks (RNN) Time Delay Networks (TDNN) Long-Short Term Memory (LSTM) Hopfield networks Boltzmann machines	
		Statistical learning	Decision trees Random Forest Support Vector Machine (SVM)	
		Probabilistic methods	Naive-Bayes Fuzzy Classifier	
		Unsupervised learning	Clustering	k-means Hierarchical clustering DBSCAN Fuzzy clustering Self-organizing map
			Dimension reduction	Autoencoder Principal component analysis
			Probabilistic methods	Fuzzy k-means
		Partially supervised learning	Statistical approaches	Expected Value Maximization (EM) with generative mix models Transductive Support Vector Machines
			Modified learning strategies	Self-training Co-training
			Graph-based approaches	Graph-based approaches
	Reinforcement learning	Temporal Difference Learning	Q-Learning SARSA	
		Monte-Carlo methods	Markov Chain Monte Carlo	
		Adaptive dynamic programming	Active and passive adaptive dynamic programming	
HYBRID LEARNING METHODS	Hybrid neural systems	Unified Neural Architectures	Constructivist Machine Learning	
		Transformation Architectures	Rule extraction for neural networks, neuro-fuzzy expert systems	
		Hybrid Modular Architectures		
	Learning via knowledge structures	Logical learning	Current best learning	
		Inductive logical programming	Sequential covering algorithm, constructive induction algorithms	
		Explanation-based learning		
		Learning using relevant information		
Conversational learning	Active, dialogue-based learning	Dialogue-based supervised learning Dialogue-based reinforcement learning		

#### 4.1.2.1.2 Classification of AI capabilities

AI as a scientific discipline is inspired by human cognitive capabilities [31]. Such capabilities have been classified within didactics and pedagogy since the middle of the last century on the basis of so-called learning goals. The most widespread classification system in use today distinguishes human capabilities both in terms of six cognitive levels and four basic cognitive domains [37], which can be used to distinguish up to 24 human cognitive capabilities.

Against this background, all currently existing AI-based systems represent only a part of the human cognitive capability spectrum. If one follows the assumption that AI capabilities imitate human capabilities, they can be roughly divided into the core areas of perception, understanding, action and communication. Most of these capabilities are realized by combining mechatronic and software components. The proposed classification helps to structure the discussion, but is not selective.

AI capabilities from the field of perception include information processing through the sensory abilities of image understanding, sound interpretation, haptics, smell and taste processing up to the complex field of recognition and interpretation of social signals.

The capability to understand is used to describe information processing in terms of evaluation, prediction and decision-making. The spectrum includes the sub-items fusion of perceptions, episodic memory, explanation and self-regulation.

The AI capability action describes in particular mechanically or physically executed activities such as robot perception, motion planning, sensor technology 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.

Communication refers to the transmission of information for processing natural language and during human-machine interaction. In computational linguistics, natural language processing corresponds to the skills of text generation, machine translation, text analysis, information and knowledge extraction, information retrieval, document analysis and speech dialogue systems. Human-machine interaction involves cognitive systems and interaction paradigms and modalities.

**Table 3:** Classification of AI capabilities

CAPABILITIES OF ARTIFICIAL INTELLIGENCE		EXAMPLES
PERCEPTION	Sensor data processing and interpretation	Image understanding Image analysis, object recognition, video analysis, perceptual reasoning, scene analysis, photometry, physical attributes, 3D modelling, simulation, virtual reality
	Noise interpretation	Noise interpretation Language recognition and synthesis, noise recognition and synthesis, anomaly recognition
	Haptics	Haptics Near-sensor technologies and methods of perception for tactile input and output (sensations like structure, tickling, touch, movement, vibration, temperature, pressure and tension)
	Social signals	Social signals Recognition and interpretation of gestures, facial expressions, body posture, affects and mood, emotions
	Smell and taste	Smell and taste Near-sensor technologies and methods of perception to recognize and synthesize smells, recognition of smell anomalies, and recognizing taste

CAPABILITIES OF ARTIFICIAL INTELLIGENCE		EXAMPLES
UNDERSTANDING	remembering, deciding and prediction	Fusion of perceptions Sensor data fusion and interpretation at the semantic level, data association, decision fusion, status assessment, ML-based/model-supported/factorgraph-based/probabilistic sensor data fusion methods
		Memories and models Episodic and semantic memory, task and process modelling, environment modelling, process memory, discourse memory, plan library
		Explanation Explanation derivation and generation, rationalization, hybrid models, integrated prediction and explanation models, explanation through architecture modification, model-diagnostic explanation
		Self-regulation Modelling own performance limits, resource-adaptive action planning, methods of self-optimization, dynamic “world modelling”
ACTION	Robotics	Robot perception Near-sensor technologies and methods of perception in robot systems, 2D and 3D perception methods, localization
	Software robots	Movement planning Methods of planning unsure movements, control methods
		Sensors and manipulators Passive and active sensors, effectors, manipulators, cooperating manipulation
		Kinematics and dynamics (movement) Kinematics systems, spatial kinematics, forwards kinematics, inverse kinematics, dynamic movement systems
		Human-robot interaction Soft robotics, human-robot collaboration, multi-modal teleoperation
		Software agents “Autonomous software systems, process automatization, (Chat-)Bots that carry out transactions, acting assistance systems”
COMMUNICATION	Processing natural speech	Text generation Paraphrasing, Markov text generation, meaning-text model, generation of relationships, reports, artistic texts
		Machine translation Transfer and interlingual methods, example-based, static, neural and semi-automatic approaches
		Text analysis Parsing (syntactic analysis), shallow and deep analysis (semantic interpretation)
		Information and knowledge extraction Text and web mining, entity extraction, disambiguation, relation extraction, event extraction
		Information retrieval Vector space model, LSA, pLSA, semantic search, fact search, question-answer systems, autocomplete
		Document analysis OCR, ICSR, document classification, segmentation, range recognition
		Speech dialogue systems Speech act recognition, reference resolution, explanation dialogue, discourse modelling, dialogue management, language change strategies
	Human-machine interaction	Cognitive systems Human factors, human processor models, user modelling, cognition theory (cognition, mental models, memory, learning type, cognitive load)
	Interaction paradigms and modalities Interaction design, patterns, multimodal interaction, user experience, fusion and fission of modalities	



Using the classification matrix for methods and capabilities, a labelling requirement for implemented methods and capabilities can be established for AI applications. [Chapter 4.3](#)

provides an overview of requirements and challenges regarding the conformity assessment and quality assessment of AI-based systems.

**Table 4:** Method-capability matrix

(CORE) METHOD-CAPABILITY MATRIX OF ARTIFICIAL INTELLIGENCE			CAPABILITIES																							
			PERCEPTION				UNDER- STANDING			ACTION				COMMUNICATION												
			Sensor data processing and interpretation				Evaluation, remembering, deciding and prediction			Robotics		Software robots		Processing natural speech			Human-machine interaction									
			Image understanding	Noise interpretation	Haptics	Social signals	Smell and taste	Fusion of perceptions	Memories and models	Explanation	Self-regulation	Robot perception	Movement planning	Sensors and manipulators	Kinematics and dynamics (movement)	Human-robot interaction	Software agents	Text generation	Machine translation	Text analysis	Information and knowledge extraction	Information retrieval	Document analysis	Speech dialogue	Cognitive systems	Interaction paradigms and modalities
METHODS	PROBLEM SOLVING, SEARCHING, OPTIMIZATION, PLANNING, DECISION-MAKING	Problem solving	Problem-solving agents, problem solving through searching, search strategies												To be taken from the previous columns according to the application											
	Optimization	Statistical optimization methods																								
		Bio-inspired optimization methods																								
	Planning and plan recognition	Autonomous and semi-automatic planning methods Plan Recognition Methods																								
Decision-making	Approaches for Decision Making																									

(CORE) METHOD-CAPABILITY MATRIX OF ARTIFICIAL INTELLIGENCE		CAPABILITIES																							
		PERCEPTION			UNDER-STANDING		ACTION				COMMUNICATION														
		Sensor data processing and interpretation			Evaluation, remembering, deciding and prediction		Robotics		Software robots		Processing natural speech		Human-machine interaction												
		Image understanding	Noise interpretation	Haptics	Social signals	Smell and taste	Fusion of perceptions	Memories and models	Explanation	Self-regulation	Robot perception	Movement planning	Sensors and manipulators	Kinematics and dynamics (movement)	Human-robot interaction	Software agents	Text generation	Machine translation	Text analysis	Information and knowledge extraction	Information retrieval	Document analysis	Speech dialogue	Cognitive systems	Interaction paradigms and modalities
KNOWLEDGE REPRESENTATION AND INFERENCE	Representation of knowledge	Knowledge representation languages and models																							
		Ontological engineering																							
		Knowledge graphs and semantic networks																							
		Modelling in formal logic																							
	Logical reasoning	Automatic proof methods																							
		Interactive proof methods																							
	Uncertain knowledge	Quantifying uncertainty																							
		Representation of uncertain knowledge																							
	Probabilistic reasoning	Inference in Bayesian networks																							
		Relational probability models																							
	Time and uncertainty in probabilistic reasoning																								
Non-probabilistic approaches	Qualitative approaches																								
	Rule-based approaches																								
	Reasoning with vagueness																								
	Reasoning with belief function																								
	Further approaches to uncertain reasoning																								
MACHINE LEARNING	Supervised learning	Neural networks																							
		Statistical learning																							
		Probabilistic methods																							
	Unsupervised learning	Clustering																							
		Dimension reduction																							
		Probabilistic methods																							
	Partially supervised learning	Statistical approaches																							
		Modified learning strategies																							
	Graph-based approaches																								
Reinforcement learning	Temporal Difference Learning																								
	Monte Carlo methods																								
	Adaptive dynamic programming																								

(CORE) METHOD-CAPABILITY MATRIX OF ARTIFICIAL INTELLIGENCE		CAPABILITIES																								
		PERCEPTION		UNDER- STANDING		ACTION				COMMUNICATION																
		Sensor data processing and interpretation		Evaluation, remembering, deciding and prediction		Robotics		Software robots		Processing natural speech		Human-machine interaction														
		Image understanding	Noise interpretation	Haptics	Social signals	Smell and taste	Fusion of perceptions	Memories and models	Explanation	Self-regulation	Robot perception	Movement planning	Sensors and manipulators	Kinematics and dynamics (movement)	Human-robot interaction	Software agents	Text generation	Machine translation	Text analysis	Information and knowledge extraction	Information retrieval	Document analysis	Speech dialogue	Cognitive systems	Interaction paradigms and modalities	
HYBRID LEARNING METHODS	Hybrid neural systems	Unified Neural Architectures																								
		Transformation Architectures																								
		Hybrid Modular Architectures																								
	Learning via knowledge structures	Logical learning																								
	Inductive logical programming																									
	Explanation-based learning																									
	Learning using relevant information																									
	Conversational learning	Active, dialogue-based learning																								

KEY: ■ Method class is often used to achieve the capability □ Method class is rarely or never used

### 4.1.2.1.3 Classification of AI applications

The classification of AI applications is often based on the AI methods and AI capabilities described above. The aim of the AI application is to concretely implement mathematical methods and abstract capabilities 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 to enable innovations in business models. In addition, the typical software markets (see Table 5) are uniformly designated worldwide and are regularly monitored by independent market analysts (e.g. IDC, Gartner, Forrester, etc.), so that potential users, projects and investors are well informed about the status of capabilities.

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

Business intelligence & decision support focuses on the timely and topic-oriented creation of reports. These are designed to provide 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. This supports decisions and enables complete planning processes 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 and the calculation of forecasts. Another area is the processing of speech typically used for searching, navigation and exploration in large text bodies. When several of these functions are combined, entire business processes can be automated, often referred to as Robotic Process Automation (RPA).

Since 2012 the AI trend has accelerated considerably due to the fact that the available CPUs and GPUs (central and graphics processing units) are becoming more and more powerful

and AI methods based on artificial neural networks can be realized faster and cheaper. 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, for example simple dialogues in call centres and service centres.

To simplify the use of AI applications, typical AI applications are offered in public or private cloud environments. This allows the user to start immediately with the adaptation of the application to their own needs without having to spend a lot of time and effort on building hardware and software. Typical AI services that are offered out-of-the-box are: image recognition, video analysis, speech-to-text conversion, text-to-speech conversion, translation, text analysis, intelligent search 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

und 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 professional software development, e.g. method-based and 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. Here, an important role is played by the following characteristics, 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 5:** Overview of software markets and typical AI applications

Software markets & typical AI applications		
Software market	Typical software products	Principles
Business Intelligence & Decision Support Systems	Business Intelligence	Autonomy & Control
	Decision Support	
	Workflow systems	
	Planning Analytics	
	Constraint Based Optimization	
	Prediction Capability	Fairness
	Text Processing Platforms & Search Engines	
	Robotic Process Automation (Rule-Based)	
	Cognitive Automation (Training-Based)	
	Real-Time Processing	

Software markets & typical AI applications		
Software market	Typical software products	Principles
AI based Customer Interaction	Chatbots	Transparency
	Voicebots	
	Avatars	
	Virtual & Augmented Reality	
AI based Services consumed from Public or Private Cloud	Image Recognition	Robustness
	Video Analytics	
	Speech To Text	
	Text To Speech	
	Translation	
	Deep Learning as a Service	Security
	Knowledge Navigation	
	Knowledge Exploration	
	Intelligent Search	
	Natural Language Processing	
AI Development Environment & Tools	Build & Develop AI	Data Governance
	Train & Optimize AI	
	Run & Manage AI	
	Ethic Support Tools	

#### 4.1.2.1.4 Classification of AI autonomy

AI applications and the computer systems that implement them can have different degrees of decision autonomy [33]. For example, the Data Ethics Commission of the German Federal Government [10] distinguishes three classes of autonomy: algorithmically-based, algorithmically-driven and algorithmically-determined systems.

Algorithmically based AI applications work as pure assistance systems without autonomous decision-making authority. However, the (partial) results and (partial) information calculated by them are the basis of human decisions.

Algorithm-driven 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.

#### 4.1.2.1.5 Risk-based assessment of applications

In view of the diversity, complexity and dynamics of applications, the Data Ethics Commission sees a need for risk-based assessment. The aim is to contribute to a human-centred and value-oriented design and use of systems. Against this background, on the basis of an ethical-legal regulatory framework, specifications for transparency, explainability and traceability are planned. Special emphasis will be placed on the aspects of the scope of information rights and obligations, as well as liability by human decision-makers.

The assessment is intended to be based on a criticality pyramid. According to the pyramid, a possible occurrence of damage (e.g. human-induced and/or algorithmically determined) is to be assessed with its extent (e.g. “right to privacy, fundamental right to life and physical integrity” and “prohibition of discrimination”) for a socio-technical system. For the assessment, the involvement of all technical components (including hardware, software and training data), human actors (including developers, manufacturers, testers and users) and life cycle phases (including development, implementation, conformity assessment and application) is sought. In addition to the legislator, developers, testers and users should also be able to assess the criticality of a system using the pyramid.

The criticality pyramid (see Figure 11) has five levels (or degrees) of criticality. As the level of criticality increases, the demands on a socio-technical system to be evaluated grow. Level 1 systems: “Applications without or with only minimal potential for harm” are checked for quality requirements and are not subject to a risk-based assessment (application example: automatic purchase recommendation; anomaly detection in industrial production). A risk impact assessment should be carried out for Level 2 to Level 5 systems. Level 2 systems, “applications with a certain potential for harm”, should have disclosure requirements on transparency. In addition, investigations into misconduct are necessary, for example by analysing the input and output behaviour (application example: non-personalized, dynamic pricing; automatic processing of claims settlement). For Level 3 systems “applications with regular or considerable potential for harm”, approval procedures should be used in addition to the measures at Level 2 (application example: automatic credit allocation; fully automated logistics). Level 4 systems “applications with substantial potential for harm” should, in addition to the measures of Levels 2 and 3, fulfil further obligations for control and transparency, such as publication of algorithms and calculation parameters, as well as the creation of an interface to directly influence the system (application example: AI-based diagnostics in medicine; automated driving). Systems at Level 5, “applications with unacceptable potential for harm” shall be subject to a proportionate or

complete ban on use (application example: systems that override the presumption of innocence, or systems that have an approvingly lethal effect without human influence).

With regard to AI, the application of the criticality pyramid has revealed a further, more profound need for discussion. In the course of this, a procedure for the legal assessment and the ethical evaluation of AI applications should crystallize. This would make it possible, for example, to define the scope of basic and liability rights for an AI application. Furthermore, the significance of the criticality pyramid could be increased by including several additional dimensions, so that a possible extent of harm can be described more concretely. In addition, certification in the course of a conformity assessment should be able to demonstrate the fulfilment of requirements with regard to the potential for harm of AI applications within Levels 1 to 4. For Level 5, the demonstration of conformity is to be prohibited, since, for example, the prevention of a high level of harm cannot be ensured through certification. In conclusion, there is the greatest variety of obligations, requirements, reservations, concerns, ethical and legal implications with regard to regulation and conformity assessment certification for systems at Levels 2 to 4.

For the assessment of AI-relevant criteria, standardized conformity assessment procedures of accredited testing laboratories can be used, for example based on the ISO/IEC 17000

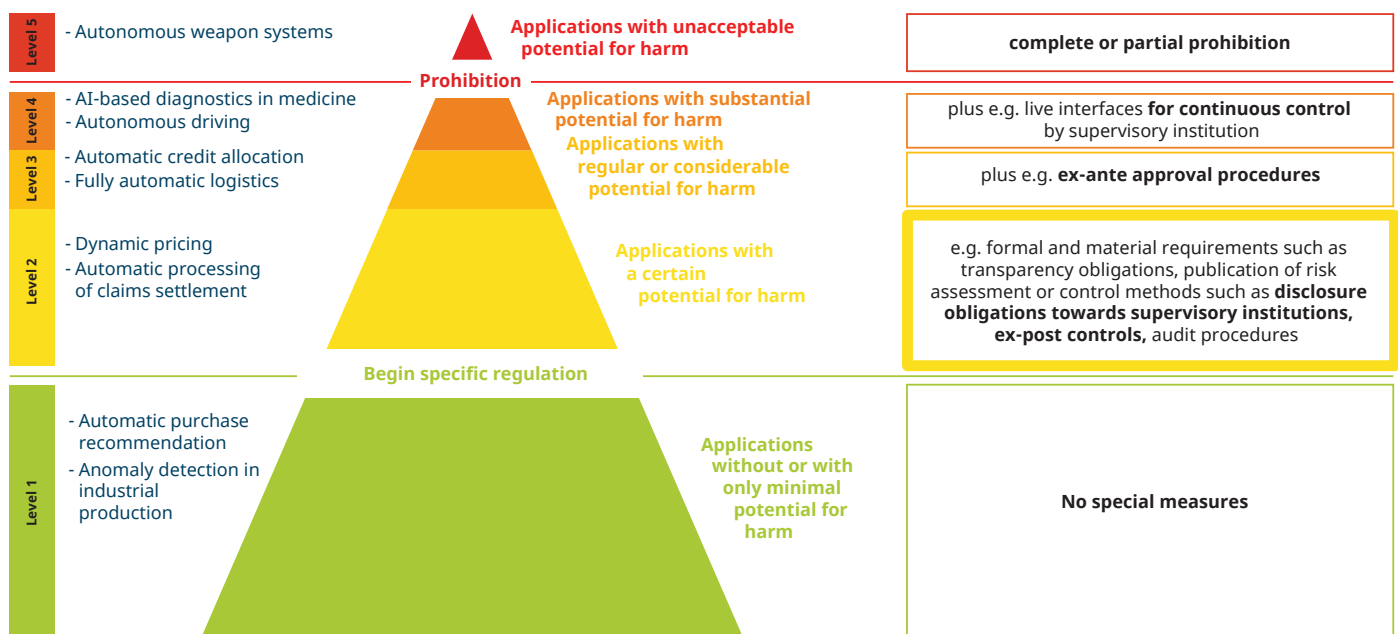


Figure 11: The criticality pyramid [10] and a risk-adapted regulatory system for the use of algorithmic systems

series of standards [38]–[44]. In the course of conformity assessment, products, systems and processes may be subject to testing, calibration, inspection or certification, and persons to certification. To this end, the expertise of already established, accredited certification bodies should be expanded with regard to the methods and capabilities of AI. An insight into relevant aspects of conformity assessment with a focus on AI is provided in Chapter 4.3.

#### 4.1.2.2 Trustworthiness

The term “trustworthiness” can basically refer to both organizations and technical systems. A technical system (i.e. a product or an electronically provided service) can be trusted with regard to certain properties such as security or reliability if there is evidence (e.g. in the form of a test report or a certificate) that the system meets such properties.<sup>17</sup> The trustworthiness of an organization is broader: It refers to an organization being trusted to implement appropriate measures and maintain management structures – a management system – to meet the expectations of its shareholders and other interested parties. In addition to a corresponding test report, the reputation of an organization or its acceptance in the market can also contribute to its trustworthiness<sup>17</sup>.

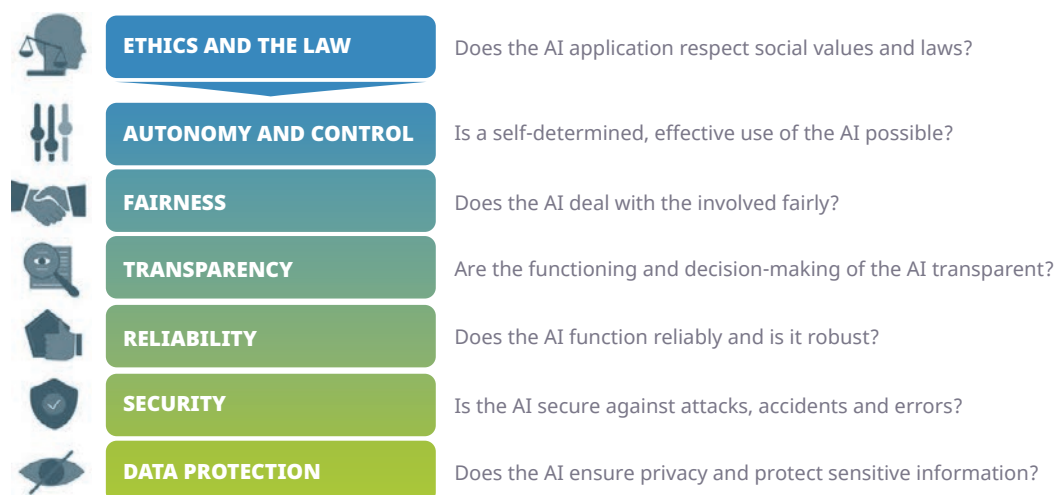
In the context of this paper, technical systems that implement AI functions (called AI systems), or organizations that implement, offer or operate such systems will be considered.

##### 4.1.2.2.1 Requirements for trustworthiness

In its ethical guidelines [5] the High-Level Expert Group on Artificial Intelligence (AI HLEG) has described a number of requirements for AI systems with regard to their trustworthiness. In most cases, these are hybrid applications, i.e. they consist of AI components and not AI-based software and hardware, and are basically understood as special IT. In this chapter, these requirements will be used as representatives for a number of similar approaches to derive standardization needs. Figure 12 gives an overview of the requirements mentioned in the guidelines, which are further discussed below:

1. **Priority is given to human agency and oversight**, and the observance and safeguarding of fundamental rights are also mentioned. It is required that information, supervision and control mechanisms are available in connection with AI systems in order to avoid negative effects, e.g. on basic rights, but also the misuse of AI systems. On the one hand, these questions have technical impli-

**Figure 12:** Requirements for a trustworthy AI [along the lines of [45]]



<sup>17</sup> Ultimately, here the trustworthiness of a technical system is attributed to the trustworthiness of an organization, namely the testing body. However, since the audit refers to the system and not to its manufacturer or provider, this distinction between system trustworthiness and organizational trustworthiness should be maintained to better structure the discussion.

cations that relate to the development of AI systems, namely the implementation of effective monitoring and control functions. However, the use of such functions must be embedded in the management processes of the operating organization in order to be effective. After all, the question of the process of human action and control of technical systems by humans refers to the objectives, the mission and the willingness to take risks of an organization operating AI systems (governance). In the context of public security, for example, different considerations will play a role for the use of AI than for use by a commercial enterprise. The AI HLEG demands that an impact assessment be carried out in areas where the use of AI may affect fundamental rights.

2. **Technical robustness and safety**, e.g. resilience to attacks and security breaches, fall back plan and general safety, accuracy, reliability and reproducibility. From the perspective of standardization, an entire range of relevant questions arise:
  - Are common approaches to management IT or cybersecurity sufficient for the use of AI? What are the specific vulnerabilities of AI systems? Are new controls or management processes necessary?
  - What restrictions must an AI system be subject to? When does the AI have to be restricted or overruled by classical systems or by humans in order to avoid damage to persons or objects?
  - How can the precision of AI systems and their reliability be measured or ensured? Are there generally accepted metrics and units of measurement? What role do development and quality assurance processes play?
1. **Privacy and data governance**, such as respect for privacy, data quality and integrity, and data access. Questions concerning standardization activities are data protection management in connection with AI, but also how data quality can be ensured overall. This applies in particular to the case where data for machine learning is provided by external providers.
2. **Transparency**, e.g. traceability, explainability and communication. On the one hand, the AI HLEG requires that data records and processes that led to the decision of the AI system be documented. On the other hand, the term “explainability” refers to the traceability of the internal function of AI systems (e.g., the question with which criteria an automatic decision was made by an AI system).
3. **Diversity, non-discrimination and fairness**, e.g. avoiding unfair bias, accessibility and universal design and stakeholder participation.

4. **Societal and environmental well-being**, e.g. sustainability and environmental friendliness, social impact, society and democracy.
5. **Accountability**, e.g. auditability, minimizing and reporting negative impacts, trade-offs and redress.

In summary, let it be said that the AI HLEG recommendations address a number of important issues. However, the publication cannot be used directly to derive mandates to the standardization bodies:

1. Standards are basically of a technical nature, i.e. they refer to requirements and recommendations of a technical-organizational nature and how such can be applied within an organization. Social, legal and political requirements cannot be codified in standards, only technical-organizational implications resulting from such requirements can become the subject of a standard. Thus, not all topics mentioned by the AI HLEG are already suitable for standardization.
2. The AI HLEG does not distinguish between trust in the AI product or service (in the sense of a product or service that uses AI functions), and trust in the organization that provides such a service or uses, manufactures or distributes such a product.
3. If standardization is seen as an objective at international level, i.e. within ISO, IEC or the ITU, an ethical basis for such work must be dispensed with unless it is generally accepted in the international community. For example, the project to propagate a framework of values that is not internationally recognized with the help of an international standard is excluded by the principles of the World Trade Organization that are binding for these three organizations [46].

#### 4.1.2.2.2 Trust in products and services

##### Common Criteria (CC)

The Common Criteria (CC) [47] describe a methodology for testing products and services with a focus on their security, which can be used as a conceptual framework for corresponding tests of AI systems. The CC are also available as an International Standard ISO/IEC 15408 [48]–[50]. A coordinated methodology for evaluation based on the CC is described in the International Standard ISO/IEC 18045 [51]. These documents form the technical basis of the Common Criteria Recognition Arrangement (CCRA) [52], which has been signed by a large number of countries, including Germany. Further information on the CC can be found on the website of the Federal Office for Information Security (BSI) [53], for example.



Requirements for testing according to the CC are summarized in the Evaluation Assurance Levels (EALs):

- 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

Certification up to EAL4 is internationally recognized.

### 4.1.2.2.3 Trust in organizations

#### The relationship between governance, management and technical-organizational measures – management systems

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 [54] (see Figure 13). It should be noted that the term “management system” refers to all three levels discussed in the following, namely the governing body, the management, and concrete technical and organizational measures.

#### Governance

Governance refers to the general tasks and the objective of an organization, its self-image and the resulting values, and the culture of the organization that determines its actions. A central concept is that of a willingness to take risks. According to ISO/IEC 38500 [54] the governing body of an organization is

responsible for the implementation of its accountability and due diligence obligations. Questions of liability are of particular relevance in connection with AI, since the possible degree of autonomy of AI raises the question of who is liable for errors and damages. Governance should take this into account, since the legal framework in this field is developing dynamically. The governing body sets requirements and establishes guidelines that must be implemented within the organization. The governing body is also responsible for establishing management structures (processes, roles, responsibilities) and providing adequate resources.

#### Management

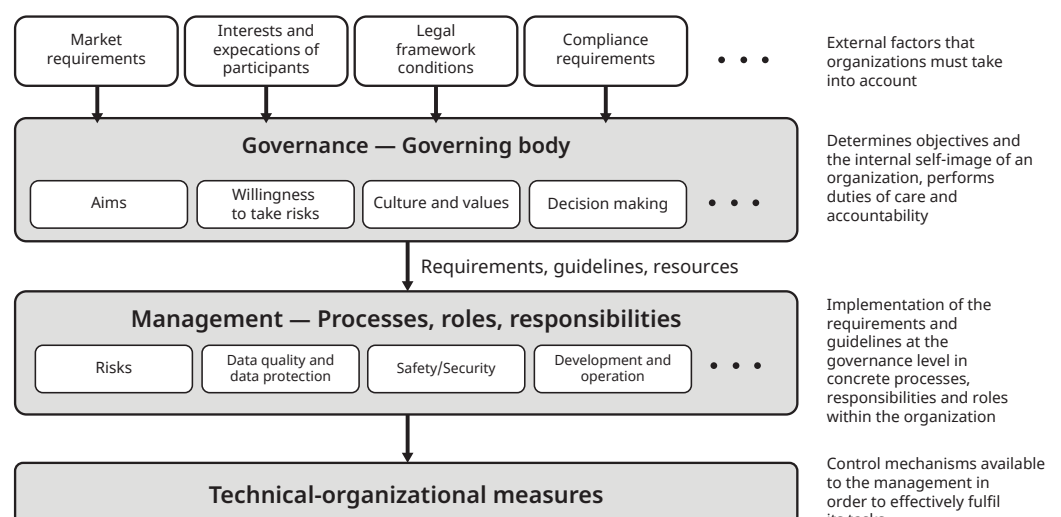
The management of an organization translates the requirements and guidelines of the governing body into concrete processes, roles and responsibilities. Examples of management tasks include:

- 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 their tasks effectively and verifiably. Technical-organizational measures range from the availability of encryption functions to increase data security

**Figure 13:** Management system: Governance, management and technical-organizational measures



to the application of statistical methods to identify unfair distortions or contamination in data sets and the availability of test and validation tools.

### Requirements on the management system

The term 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 governing 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 13](#). Minimum requirements for the management system are described in the Guidelines for International Standardization of ISO/IEC, in the so-called: High Level Structure (HLS) [55]:

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 governing body must define binding readiness of the organization and lay it down in the form of guidelines. It must also define processes, roles and responsibilities for effective management.
3. **Planning;** this must describe activities to deal with risks and opportunities.
4. **Support;** this includes the provision of resources, the determination of necessary competencies, ensuring necessary mindfulness, communication and documentation.
5. **Operation;** this is the operational implementation of management requirements.
6. **Performance evaluation;** this comprises monitoring, analysis and evaluation, internal auditing and management review.
7. **Improvement;** this deals with the identification of non-conformity with regard to MSS requirements, corrective measures and the continuous 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 the 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 adding AI-specific requirements to existing MSS.

However, since the different MSS are published and maintained by different bodies in ISO and IEC, which have neither a common conceptual framework nor a synchronized way of working, and since it is not clear whether existing MSS are even sufficient to cover all aspects of AI, it is more promising to design a new MSS that focuses on AI-specific requirements.

### Supporting specifications

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 now contain no 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 and there are methods, such as V-Model® XT [56], [57] – with and without agile methods (e.g. Scrum) – which help ensure the quality of the overall result during development. For software with a predetermined functional sequence, there are generally accepted development and quality assurance procedures, such as code reading, module and application tests at various integration levels, verification and validation. These methods and procedures also work for 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 also required. This applies in particular to evidence when proving functional safety in product liability.

An appropriate approach to the development of AI systems is a risk-based approach<sup>18</sup> considering the entire life cycle of an AI system in its application environment, as well as ensuring data quality in the learning and application phase.

Further consideration must be given to AI systems whose source code and/or learning content was generated by themselves or by other AI systems. Thus, an existing AI system develops a new one or changes its learning content, so that a kind of evolution of the machines takes place.

#### 4.1.2.3.1 The life cycle of an AI system

Similar to traditional software development, the life cycle phases of an AI system consist of: **Concept, Development, Deployment, Operations and Retirement**, whereby especially for systems based on machine learning, which can be applied in different phases of the life cycle from development to operation, there is a much closer interlocking of the phases than is the case with classical software systems.

During the concept phase it has to be defined whether the application to be created is created as a rule-based, static or dynamic AI module and which requirements result from the context of the application area, as well as the necessary data quality. For rule-based AI systems, the established software life cycle according to ISO/IEC/IEEE 12207 [58], or for safety-critical systems also according to ISO 26262 [59]–[70], ISO/IEC 27034 [71]–[78] or IEC 61508 [79]–[86], can be applied. A risk-based approach is necessary for static and dynamic AI systems.

18 “Risk-based” in English.

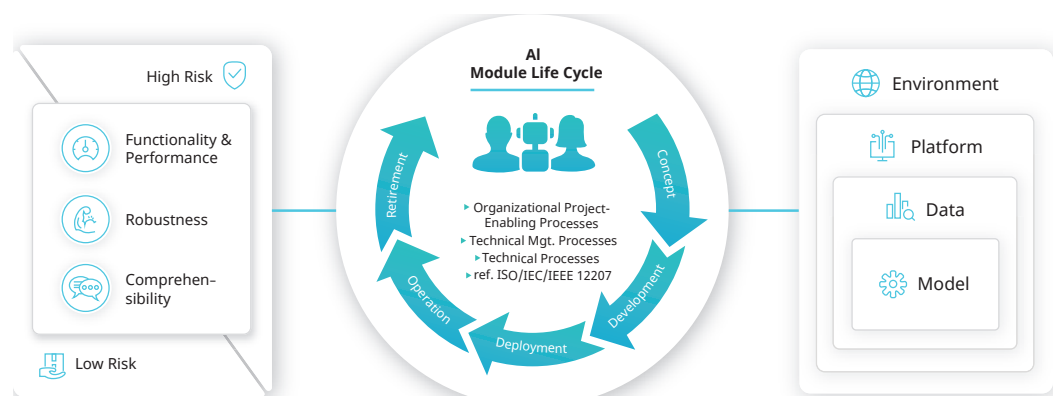
Based on this, a risk analysis must be carried out, e.g. based on an FMECA (Failure Mode and Effects and Criticality Analysis), which must consider the entire life cycle of the AI system. As part of the risk assessment, a first simple classification, as presented in DIN SPEC 92001-1 [87], can be made. (see Figure 14). The separation into low risk and high risk can be sufficient, but a more fine-grained phase model seems to be more appropriate, especially since aspects of dynamic models can be dealt with in more detail.

As an alternative to the previously described DIN SPEC 92001-1 [87], VDE/DKE presented a “Reference Model AI” [88] (see Figure 15) which describes a development process for AI systems based on the V-Model® XT. A consensus model for the AI life cycle is to be developed within the framework of the Standardization Roadmap AI.

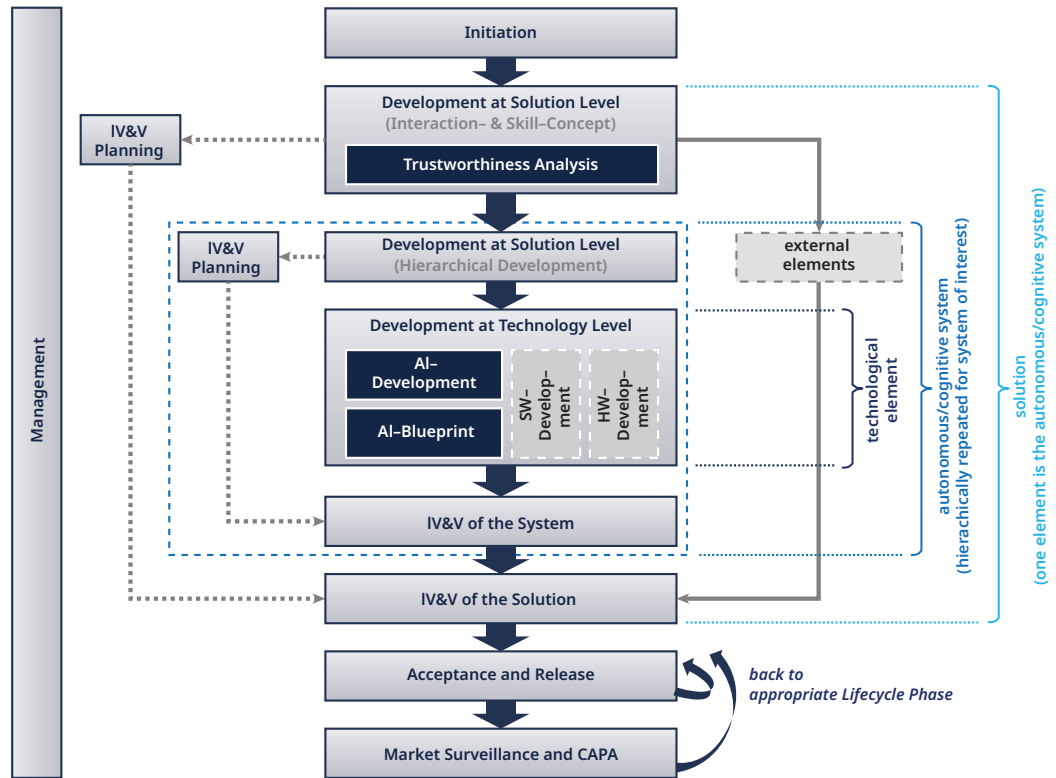
#### 4.1.2.3.2 Data quality principles for AI modules

The quality of the data for learning, testing and subsequent application is an essential factor for the successful development and, in the application phase, for the use of AI systems. A general definition of data quality in software development is described in ISO/IEC 25012:2008 [89] and consists of inherent and system-dependent characteristics. To what extent this standard is also suitable for the development of AI applications, or if other or further quality features are important, has to be checked and, where necessary, standardized specifically for AI applications. It will make sense to tailor and/or prioritize the dimensions for the respective use case. If, for example, simulation data (“synthetic data”) are used for learning and/or testing, their usability/exemplary nature must be ensured. If incorrect data is deliberately provided for learning, testing and inspection purposes, it must be marked

**Figure 14:** AI quality meta-model of DIN SPEC 92001-1 [87]



**Figure 15: VDE/DKE Reference model AI [88]**



accordingly and separated from the non-erroneous data in a suitable manner so that no unintentional mixing occurs.

The Fraunhofer Guidelines for High Quality Data and Metadata (NQDM) of 2019 [90] lists the following dimensions of data quality:

1. **Currency**

Data describe the current reality. Therefore, it is recommended to pay attention to a time stamp and, if necessary, a version number when recording and naming the data. Data should be checked at appropriate intervals to ensure that they are representative.

2. **Accuracy**

The data should contain correct values and be as error-free as possible. Here a datum is faulty if it does not correspond to its classification. Thus an incorrect datum, which has been communicated to the AI system for training as incorrect, is not incorrect in this sense. For the training of AI systems, incorrect data is deliberately used, but it is also classified as faulty.

3. **Precision**

Depending on the application, the precision of the data is of high relevance, so that, for example, rounding of values should be avoided. The content descriptions of the data should also be as precise as possible in order to quickly assess the relevance of data.

4. **Conformity**

When providing data, attention must be paid to the expectation conformity of the contained information in a certain usage context and format, for example when naming attributes and vocabulary. For a universal use of the data, appropriate standards should be used where possible, e.g. ISO 8601 [91] for dates.

5. **Consistency**

Data should be free of contradictions, both in itself and across data sets. This dimension may already be covered by accuracy.

6. **Transparency and trustworthiness**

The origin, originality and changes to the data should be made traceable, so that the transparency and credibility of the data can be strengthened, thereby gaining the trust of the users and also meeting ethical requirements.

7. **Reliability**

In order to assess the reliability, or the degree of maturity, of a piece of information, it can be assigned a status (see also [DCAT-AP.de](https://www.dcat-ap.de/)).

8. **Understandability**

The data structure, the naming of the data, as well as data interfaces should be easy to understand.

9. **Completeness**

A data set should be complete: Attributes, which are mandatory for the further use of the data set, must therefore contain a value.

#### 10. Accessibility and availability

The resources should be easily accessible. This includes easy findability, long-lasting links and references, as well as comprehensible descriptions.

To achieve a high level of data quality, a precise specification of the requirements for data and data interfaces is necessary. Results from the Platform Learning Systems show that data management can be seen as the foundation for learning systems [92]. For the trustworthiness and traceability of applications, as well as for the assessment of their quality, a deep understanding of all individual components of the data science process is necessary, as well as for the process as a whole. Among the components of the process are: data acquisition, data cleansing, data integration, data exploration, data analysis, modelling, data visualization and data interpretation, as well as interactive processes or feedback loops within the entire process chain (e.g. monitoring, evaluation).

#### 4.1.3 Standardization needs

##### NEED 1:

##### Support of international standardization work on an MSS for AI

A project to develop an MSS for AI was recently initiated in ISO/IEC/JTC 1/SC 42 “Artificial Intelligence” [Note: Actually, the proposal is currently being voted on, but a positive vote of the national representatives in SC 42 can be considered certain]. Since such a standard is ultimately fully certifiable and will thus represent an International Standard for requirements and processes for organizations developing or using AI, participation of German stakeholders in this project is strongly recommended. Implementation activities within the framework of the Standardization Roadmap AI should in particular consider providing funds and resources for such participation.

##### NEED 2:

##### Drawing up of a technology roadmap for AI

In addition to the AI classification methodology outlined above in 4.1.2.1, 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 3:

##### Risk management for AI

Based on the International Standard ISO 31000 Risk management [93] a project on risk management for AI is currently being carried out in ISO/IEC JTC 1/SC 42 under the number ISO/IEC 23894. In its current version, the document describes extensions of the generic guidelines from ISO 31000 for AI-specific aspects. Risk management must continue to be complemented by impact assessment guidelines for the use of AI systems.

##### NEED 4:

##### Data quality management for AI

Data quality management is a priority issue in the context of machine learning. A number of data quality management projects are currently being initiated in ISO/IEC JTC 1/SC 42 and are expected to start in autumn 2020, which should be critically observed by German participants and supported by contributions if necessary.

##### NEED 5:

##### Management of transparency and avoidance of discrimination

As mentioned in 4.1.2.2, the explainability and traceability of AI systems is another topic related to AI, which should be the subject of standardization. This should be supplemented by the definition of technical and organizational measures to prevent discrimination.

##### NEED 6:

##### Design principles for KI systems

Work on the definition of a life cycle model for AI systems is currently already being carried out nationally within the framework of DIN SPEC 92001 and internationally in ISO/IEC JTC 1/SC 42. These activities should be harmonized and continued within the framework of International Standards.



## 4.2

## Ethics/Responsible AI



Ethics is a special field of philosophy and the basis for the responsible use of technology in general and AI in particular. A short excursion on the basics of philosophy and thus its special field of ethics in our culture, the terms of the ethical dilemma, and AI-ethics is given in the [Annex 11.2](#).

Responsible AI is about creating a framework for the assessment, deployment and monitoring of AI to create new opportunities for better services to citizens and institutions. It means designing and implementing solutions that focus on people. Using design-oriented thinking, organizations examine core ethical issues in context, evaluate the appropriateness of policies and programs, and create a set of value-based requirements for AI solutions.

Algorithmic decision systems, in particular those that derive their decision rules from historical training data using machine learning methods, arise in a long chain along which responsibilities are distributed. This system concept explicitly includes people and processes. The responsibilities have to become the focus of work in the course of standardization efforts in the field of artificial intelligence. Standardization can help to make the handover points in this chain of responsibility transparent, thereby enabling modularization that allows specialists to find the best solutions in competition.

It should be noted that generating AIs with ethically relevant aspects, such as deep fake technology (imitation of people and their behaviour in images, including video and sound), are not considered. Although they do have far-reaching ethical issues due to their possibilities, these relate solely to the application of these systems and less to the development and creation process that standardization work will have to deal with in the coming years.

It is generally established that the exclusive consideration of the technical component is not sufficient. The possible use of the same technical component, i.e. the same decision system in different fields of application, clearly shows that a certification of the purely technical parts cannot do justice to the complexity of the problem. A socioinformatic overall view is therefore appropriate, taking into account all social actors, as well as the embedding of the automated decision-making system (ADM system) in the social process. However, since ADM systems are always subject to the legal framework, which is currently undergoing major changes and adaptations (see Directive 2006/42/EC (Machinery Directive) [94], Regulation (EU) 2016/679 (General Data Protection Regulation (GDPR) [95], etc.) bridging the gap between standardization

efforts and legal conformity will be a complex task of future standardization work (e.g. questions of hazard prevention and liability issues). It must always be taken into account that the infringement of legal rights when using AI is often difficult to detect and prove.

#### 4.2.1 Status quo

In this sense, large initiatives, political statements and expert commissions dealing with ethics principles, values and criteria which have been laid down in policy and position papers, reports and studies have emerged in parallel to AI systems that are penetrating more and more into all areas of life, especially through the increasingly elaborate systems of machine learning (ML). These in turn are supported and initiated by political strategies of the EU and individual nations. Especially in the European context, the need for a sovereign handling of the advancing digitalization is being addressed in the course of AI development. The keyword digital sovereignty is used to describe, among other things, the empowerment of citizens and the design of human-centred offers.

Most of these initiatives have an interdisciplinary perspective, consider several fields of application, and/or explain the broad background knowledge of the experts involved. This results in a varied collection of statements, values and criteria.

In addition, the topic is being actively deepened in research through calls for proposals from foundations, the German Federal Government and/or the EU through a large number of projects (e.g. of the German Ministries of Labour and Social Affairs (BMAS), of Education and Research (BMBF) and of the Interior, Building and Community (BMI)).

Interdisciplinarity in particular requires the importance of a jointly developed understanding of concepts, including common agreement on the vocabulary used. This is implemented here in the Glossary (see [11.1](#))

#### 4.2.2 Requirements, challenges

The current discourse about ethics and AI is dominated by two thoughts, on the one hand opportunities and potentials are discussed; on the other hand ethical requirements are understood as red-taping, which slow down economy and society and prevent AI systems (on the worldwide market) from



being used unhindered/economically. This concern is not unfounded, since the use of AI-based ADM systems poses some additional ethical challenges compared to other algorithms. However, it is possible to ensure compliance with minimum ethical standards through standardization.

Systems that are, or were already, in use have shown that the consequences of their use are sometimes difficult, if not impossible, to assess in advance. This can be seen very clearly in the way the American Civil Liberties Union (ACLU) handles the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) system, which is used in the American justice system to predict the probability of offenders' recidivism [96]. In 2011, the ACLU called for the use of algorithm-based decision-making systems in all phases of the penal system in America [97] with the argument of a more objective decision for defendants and convicts. Following several studies on the lack of fairness or non-discrimination of such systems, the ACLU joined a contrary demand in 2018 and argued for a ban on learned ADM systems in a judicial context [98].

Especially from the field of the assessment of technological consequences, we know that the consequences of a system can only be estimated in sufficient detail for it to be adapted accordingly if the technology is used by a sufficiently large number of people over a sufficiently long period of time. By then, however, the system is already so established that the implementation of fundamental changes is very difficult, perhaps even impossible. This problem is known as the Collingridge Dilemma [99]. For this reason, the development of new technologies and new applications constantly confronts legislators with new challenges, which they usually have to meet under great time pressure. Here standardization can help by defining and describing interfaces and review bodies, and thus supporting a reviewing body in such a way that it is sufficiently flexible to counteract the far-reaching ethical consequences that may arise in this context. This offers the developing industry a certain degree of adaptability to a constantly changing catalogue of requirements.

In essence, AI systems are mathematical components which serve optimization and/or classification and are embedded by further program elements into decision situations or (partial) automations relevant for humans. Nevertheless, an “anthropomorphization of AI” can often be observed in AI contexts, i.e. also in the context of dealing with AI ethics [100]. Human characteristics and behaviour – such as thinking, learning, decision-making, etc. – are attributed to AI. Anthro-

pomorphizing AI can be detrimental to communication on the subject, as it can obscure technical facts and inappropriately emotionalize problem references. For this reason, it is of utmost importance when dealing with AI to clearly work out the system boundaries and, above all, to separate the behaviour of technical AI components from human behaviour. To make matters worse, some systems continue to learn in use, so that a single test at a given time may potentially be insufficient for the lifetime of a system.

In the following, the approaches to “decision-making AI” that are in the main focus of this Roadmap will be distinguished from approaches to “generating AI” that are not primarily in the focus of this document. A clear, fundamental separation is impossible; however, practice shows that artificial intelligence approaches can usually be divided into one of two categories: approaches whose essential function is to make an abstracting or complexity reducing “decision” or “assessment” from input data (“decision-making AI”), and approaches that generate fundamentally new data, and either do not depend on input data at all, or are intended to significantly increase the complexity of the input data by synthetic additions (“generating AI”).

It should first be noted that this distinction has no relation to the complexity of the AI: There are simple decision procedures (e.g. Bayes classifiers) and simple generation procedures (e.g. cellular automata like Conway's Game of Life), but also complex decision procedures (e.g. object recognition using neural networks) and generation procedures (e.g. generation of photorealistic faces using generative neural networks). There are also procedures that cannot be assigned to any category, such as language translation, or procedures in which decisions can also be made on the basis of generated data (for example, through synthetically generated phantom images in police investigations).

Nevertheless, the distinction and focus on “decision-making AI” was found to be appropriate for the WG Ethics working on the Roadmap: In “decision making AIs” (to which the described challenges [97], [98] also belong), the ethical implications already arise in the conception and development phase of the systems. Here, standardization can help to accompany the development process and minimize ethical risks in application. With “generating AIs”, on the other hand, the afore-mentioned Collingridge dilemma usually occurs to a great extent: The increase in complexity results in innumerable conceivable applications from a comparatively simple procedure, on the basis of which the ethical evaluation only

becomes possible. Thus, for “generating AIs” the ethical implications are not mainly revealed in the development process, but only indirectly in the respective applications, which are often difficult to assess; these are therefore not the focus of this chapter on ethics.

**4.2.2.1 Ethically relevant problems in the development process of AI systems**

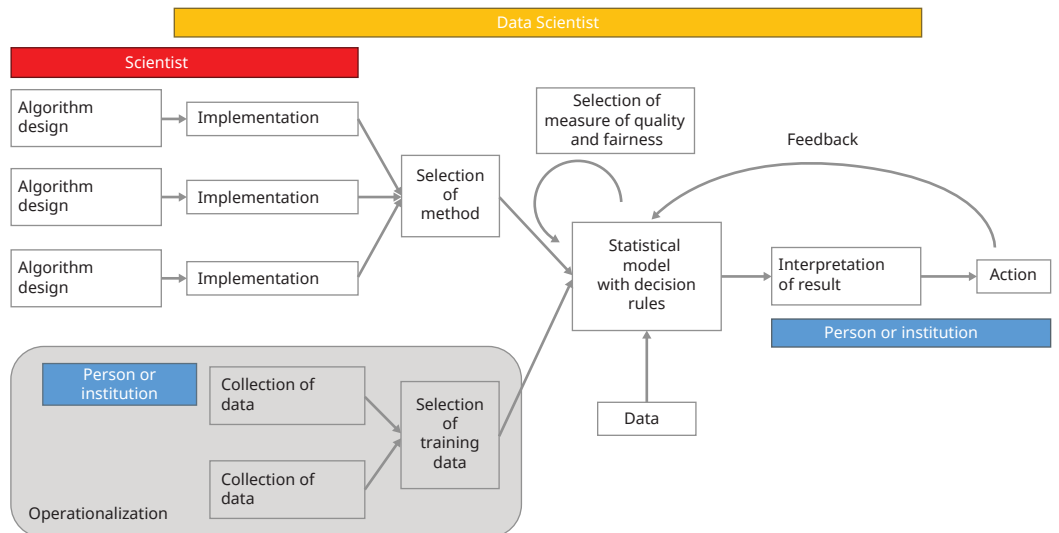
The different actors involved in the development, implementation and evaluation process have different requirements when it comes to the applicability of frameworks for ethical considerations. Providers of AI systems need approaches that make the implementation of such principles as easy as possible. This is made more difficult by the fact that large systems are often developed in delimited functional subsystems. The complexity resulting from the interaction of these subsystems makes it difficult to get an overview of the extent to which the overall system meets ethical requirements. To make this complexity manageable, it requires a process that evaluates both the subsystems and the overall system. Regardless of the actual implementation, it must be taken into account that systems that continue to learn after deployment definitely require additional support and regular evaluations. This insight is already evident, for example, in the further development of the CRISP-DM (Cross-industry standard process for data mining) industry standard to ASUM-DM (Analytics Solutions Unified Method for Data Mining/Predictive Analytics), which takes post-deployment support into account, see [101] or in the reference model for a trustworthy AI of the DKE [87].

Changing conditions in the environment of the application can lead to unexpected results, because the training phase was not focused on these conditions, which is why a consideration of the application context is absolutely necessary. In addition, it would be mandatory to perform a re-evaluation with every change or extension of the application context.

In recent years, the availability of large amounts of data paired with the technical possibilities has expanded the fields of application of AI enormously. The impact and interaction of their use in socially relevant processes has hardly been researched yet. In various contributions, the participation of stakeholders in the design of ADM systems was pointed out to ensure their acceptance [102]–[104]. Especially with regard to systems with far-reaching ethical issues, the integration of as many different stakeholders as possible in and around the development process is therefore desirable. Here the interests of the “stakeholders” [105] must be sufficiently considered. In order to grasp the complexity of the resulting decisions and decision preparations and to be able to make recommendations for action based on them, it is necessary to consider their development up to their integration into the social process, where a long chain of responsibilities is formed (see Figure 16).

**Algorithm design and implementation** The development and implementation of ADM systems is extremely complex and interspersed with many design decisions. Therefore, software packages provided by companies or communities of programmers are often used, which provide ready-to-use implementations of important components. Design decisions,

**Figure 16:** Long chain of responsibilities (as in [106])



such as the choice of some hyperparameters<sup>19</sup>, can often be intransparent or not even visible. A check whether the used software package is suitable for the application context in question often does not take place.

**Selection of methods** There are many (sub-)methods of machine learning that can be combined to a large extent. For example, an ADM system consists of at least two major components, one that learns based on the training data and one that makes a decision based on the model created by the first component. The differences not only affect functional aspects (training duration, error-proneness, ...), but also non-functional requirements that are particularly relevant in an ethical context (e.g. explainability and traceability). Each method comes with model assumptions that must be ensured by the data, the setting and the type of training in order for the method to work in a goal-oriented manner and to keep any assurances about the quality achieved. Data scientists usually lack the necessary training in the application context to recognize potential risks when using the respective technology without any explicit references. This area should play an important role in standardization efforts in the coming years.

**Data collection and selection:** If data are obtained from external sources, such as governmental, economic or scientific institutions or data vendors, there is a risk that they are or will be falsified due to an inadequate collection process or faulty preparation. The areas of data collection, data processing and data storage must be considered in detail in the context of artificial intelligence. Legal frameworks such as the GDPR can provide a rough framework, but this must be further clarified and made more precise or adapted to the application context through standardization work. It must also be taken into account that the interaction and interweaving of machine-related and personal data will also lead to new challenges. Standardization should take into account a given purpose limitation of the data, taking into account the respective field of application, which ensures that documentation for which purpose the data were collected is available. Regulation allows the establishment of conditions under which the data may or may not be transferred to another purpose. Data and method selection do not follow a fixed order, but rather different data and methods are tested in combination, as long as the necessary effort is reasonable. The goal is the operationalization of an abstract quantity that is to be made meas-

urable (e.g. creditworthiness, or relevance of a message). The results are greatly influenced by the data basis (especially selection and quality).

**Construction of the decision-making system:** In the construction of decision-making systems, training data and methods of machine learning are brought together. Whether or not the predictive power of the decision-making system meets the given requirements is determined on the basis of a self-chosen quality criterion. The choice of the quality criterion (over two dozen can be considered) is often also up to data scientists and is therefore highly subjective. The selection of the quality criterion can have far-reaching consequences, for example if an inappropriate measure is chosen, which is why this process would benefit from standardization. The system is optimized (there are many parameters that can be set) and/or trained until the requirements of the selected quality criterion (e.g. falling below a certain error rate) are met.

**Embedding in the social process:** When embedding the decision-making system in the application context, it is determined how results can be interpreted and how to deal with them. The users are instructed in some way how to use the system and can receive outputs based on their own input, thus enabling quality control in the use of the system. Often a data scientist is entrusted with the control of the quality of the system in concrete use, for example, when it comes to graphic processing. Together with the users, the data scientist can thus contribute to the interpretation of the results, on the basis of which it is decided how to proceed with the results. An automatic decision-making system can independently select and initiate actions based on the results.

**Re-evaluation:** When the development and integration process is complete, the overall system is usually evaluated again, either by the data scientist or by the users. Depending on the evaluation result, any subcomponents of the technical system can be changed again or even replaced completely (see feedback arrows in [Figure 16](#)). For example, there are known cases in America in which an AI has identified the distance to company headquarters as a relevant indicator for the probability of termination as part of a company's recruitment process. Since this discriminated against applicants who could not afford an apartment in the expensive surroundings of the company headquarters, the developers excluded this criterion from the decision-making process [107].

It should be mentioned that correctly executed agile development processes could reduce the problems in general, since

<sup>19</sup> Parameters that are defined before the actual training of an AI, such as the structure of an artificial neural network

each developer potentially also acts as a data scientist, can (alternately) take on any role in the chain of responsibilities, and there is significantly more communication between developers, which also allows individual expertise on specific aspects to be better disseminated. Nevertheless, it is only a reduction of the problem. Depending on the system, several hundred people may be involved in the development (e.g. automated driving, where the overall system is created from a large number of modules that are partly developed independently of each other, but influence each other in their behaviour with respect to the overall system). The resulting complexity can no longer be compensated by agile development methods. Furthermore, embedding into the social process is beyond the reach of software developers, which is why it can only be considered to a limited extent.

These examples show that it is important to observe the product as used, i.e. its use “in the field”. As already described above in the basics, despite evaluation, uncertainties about the behaviour of the AI system after delivery, the “residual risk”, remain in learning AI systems due to insufficient quality methods to fully verify what has been learned. Product observation in the field also includes interaction with external products that can be associated with the AI system. The aim is to systematically identify previously undetected, unwanted and unethical behaviour during the application and to take corrective action to limit or, better, to prevent unnecessary hazards or unethical behaviour. Such a product observation with feedback, as part of the AI life cycle, is described in detail in 4.3.2.3.2.4.

**Assumption of responsibility/liability:** In the development of applications of higher complexity, the early involvement of non-technical disciplines in the work steps of “data scientists” is highly advisable, taking into account the ethical perspective as well as interested parties. In this way, risks, incorrect or systematically distorted judgements, and undesired effects can be identified and minimized at an early stage. In the context of the assumption of responsibility, several evaluative process steps should already be provided for during project planning. Ethically good AI can be marketed sustainably and on a large scale. With increasing complexity of the AI, a systematic ethical reflection already in the investment costs is worthwhile. The complex interaction of many actors raises the question of who is responsible in case of damage. In a long chain of responsibilities, this can be the data scientist, the user of the system (e.g., the person who integrates ADM systems into their products and services), and the data subject (e.g. the person who uses or comes into contact with

the products and services). The point where the responsibility of one actor ends and that of another can begin is seen as problematic. Here “points of transfer”, as they are known elsewhere in our legal system, could provide more transparency. Standardization could make an important contribution to the definition and design of such “points of transfer” due to its proximity to the respective applications. A useful difference in perspective here is between the law, which clarifies questions of liability and damage regulation in the narrower sense and within strict limits, and ethics with the concept of responsibility which goes far beyond questions of the liability of individual participants and starts at the other point of the process. By clarifying responsibilities for the quality of the AI in terms of compliance with user-relevant ethical values, risks can be identified and minimized well in advance of liability issues or economic losses. Standards can require the early involvement of interested or relevant parties and can also define process chains with points of responsibility and thus potentially perform a preventive task to avoid or reduce potential risks and damage. By ensuring high ethical quality, standardization can offer a broad-based marketing potential for “good products” and help open up markets with a different legal framework.

#### 4.2.2.2 Fairness and freedom from discrimination

Although automated classification or decision-making systems can claim to avoid human prejudices and inadequacies by calculation, they are not immune for their part from producing discrimination and injustice (“systematic discrimination”, see [108]). Alongside hardware problems and problems due to the interplay between hard and software, this is primarily due to the technical logic of the systems [109]: Intelligent machines learn from available training data, which in turn is a reflection of the previous behaviour of persons and institutions, as well as the way the data was collected. The training data also reflects the mistakes or subjective assessments of people and sometimes leads to the perception of the systems as “unfair” or “discriminatory”. The perception itself is a subjective one that can neither be changed by the legislator nor by standardization. However – and here standardization can certainly make an important contribution – the collection and processing processes, i.e. the decision-making processes up to the ADM system, i.e. the procedural rules, can be made “fair” and “non-discriminatory”, (e.g. [108]). Usually five criteria are mentioned, according to which the fairness of procedural rules should be designed and also evaluated: (1)

consistency, (2) neutrality, (3) accuracy, (4) revisability and (5) representativeness (among others [110]).

- Consistency: The decision-making rules should be applied consistently, regardless of the decision-maker, the people affected and the time of the decision.
- Neutrality: The personal (process) preferences of decision-makers should not be able to influence a decision. Neutrality thus refers to the supposed impartiality of automated decision-making systems.
- Accuracy: Fair decisions should be based on the most complete and correct information possible. This addresses the reliability and validity of data input in the case of automated decisions.
- Revisability: A fair decision-making process ensures that incorrect or inappropriate decisions can be reversed.
- Representativeness means that meaningful data is available and taken into account for different identities, cultures, ethnic groups and languages that are the subject of the procedure.

Standardization can provide important impulses here, for example through specifications for decision-making processes (flowcharts) for the collection and processing of data, procedures to be followed when revising errors, design of any product monitoring obligations, etc.

#### 4.2.2.3 Canon of values

In a canon of values, which is to be considered in the development of a machine system, the values are not unrelated to each other but contain complementary and competing values and also such which are relatively independent of each other. Here it is absolutely necessary to consider the concrete application context of the ADM system, since open ethical questions in the social area are often also affected. Especially for the competing values, it can be determined that:

In practice, the fundamental evaluation of risk situations often leads to a quantification of the risk from an economic perspective. In concrete applications, no canon of values is free from an economic evaluation. This is particularly evident in the (current) Corona crisis. Although it should be undisputed that human life is the highest good against which other values can be measured, it can be observed that in many places and in different countries there is a discussion about the extent to which the economic damage caused by maintaining quarantine regulations is not greater than the damage caused by lift-

ing quarantine regulations. Which is in consequence nothing else than to set human lives against economic success. One risk is weighed against another.

#### 4.2.2.3.1 Guidance

Ethical guidelines for algorithmic decision-making systems are discussed in different contexts (e.g. [5], [32], [111], [112]). For example, there are ethical guidelines for the statistical practice of the American Statistical Association, the “Code of ethics and professional conduct” of the Association for Computing Machinery (ACM), and the ethical guidelines of the Society for Informatics (a discussion of the frameworks mentioned here can be found in Garzcaræk and Steuer [113]. A comprehensive presentation of the different AI guidelines can be found in Fjeld et al. [114], Jobin et al. [115], or Hagedorff [116]).

The report of the Data Ethics Commission [10] should also be mentioned here, which deals with general ethical and legal principles (human dignity, self-determination, privacy, security, democracy, justice and solidarity, and sustainability) in Section B. However, these are placed there in relation to ethics and law within a more general framework.

Also worth mentioning is the Ethics Briefing [117], a guide for the responsible development and application of AI by experts of the Platform Learning Systems. It contains recommendations for an ethically reflected development and application process of AI systems. These recommendations can be broken down into three fundamental values (self-determination, justice, and protection of privacy and personality) as well as further principles (promotion of autonomy, sense of responsibility, equality, freedom from discrimination, diversity and variety, fair access to the benefits of AI, sustainability, privacy as withdrawal from the public sphere, anonymity as protection of privacy, informational self-determination and integrity of personal identity) and necessary preconditions for implementation. These considerations are preceded by the three basic assumptions avoidance of damage, legal conformity and technical robustness [117]. In addition, another White Paper presents criteria for successful human-machine interaction. This can be divided into four clusters: protection of the individual, trustworthiness, sensible division of labour, and favourable working conditions [118].

### 4.2.2.3.2 Values

There are different approaches to establishing a canon of values for the selection of options for action or for the evaluation and thus also for the assessment of the risk of ADM systems. Implicitly, however, it always boils down to determining firstly the (target) values of a canon of values, secondly the relationship between these values (relationship relations), and thirdly how values and relationship relations change for the respective area of application of the ADM system. Examples for a certain approach are the basic values discussed in the already- mentioned report of the Data Ethics Committee [10], the ethics briefing of the Platform Learning Systems [117] and the criteria for successful human-machine interaction of the Platform Learning Systems [118], as well as those from the White Paper “Ethical aspects in standardization for AI”.

The definition of a concrete canon of values can become difficult in a given application context. Therefore, a methodical and systematic consideration of the individual values and their (relational) relationships is indispensable (see Operationalization of values). Thus, in a specific application context, some of the defined values may only have a marginal effect or have already been sufficiently taken into account by existing laws. In any case, it is expedient to define operationalizable values that allow a concrete measurement of the respective risk potential that a concrete ADM system represents for humans in the respective context. The definition of operationalizable values also favours their justifiability in an international context, even if concrete value concepts differ. A differentiated consideration of values depending on the concrete field of application (e.g. medicine, mobility, ...) can be helpful. The considered values often do not exist independently from each other, but are related to each other. Depending on the context of use, they may gain or lose weight. The following examples from the environment of autonomous<sup>20</sup> machines shall illustrate this:

1 Examples of software restrictions: A vehicle automatically makes an additional speed reduction in areas with increased hazard potential (e.g. in front of nursery schools). A machine in the production process reacts to the approach of a person with an immediate stop.

20 The term “autonomous” here means – in accordance with the definition in 4.1.2.1 – for vehicles/machines that they grasp their environment and move without human intervention in the same traffic space as humans. They are of course subject to hardware and software restrictions.

2 Examples of physical restrictions: Production machines in which a barrier prevents people from being injured. For autonomous vehicles it would be the maximum speed, which cannot be exceeded due to the design.

These examples make it clear that the values „human autonomy“ and „safety“ are weighted differently depending on the circumstances on the one hand, while on the other hand the importance of other values approaches zero, e.g. the transparency of communication mentioned in the results paper on the Ethics Roadmap.

This shows that in a concrete application the weighting of different ethical requirements does not necessarily have to be constant, in fact, it usually will not be. The reason for this lies in the application case and not in the ethical values. This value always depends on the specific context and circumstances.

One possibility to set up such values is to name examples of application contexts (like the case studies mentioned here) and to set the given attributes in relation to each other for the respective concrete application context. This can be done on the one hand with the help of specialists in the respective field of application, e.g. in medicine, aircraft construction, etc., and on the other hand through public participation. In addition, an increase in subject-specific and social acceptance could be expected (see [119]).

All in all, it can be seen that a fixed canon of values can be insufficient in a concrete application. In order to evaluate a concrete situation, a method that is widely used in industry and also in ISO Standards, the risk-based approach<sup>21</sup>, offers itself.

For the design of a specific canon of values for an AI system, it will not always be possible to estimate the risk of an option for action with sufficient accuracy and thus the effect on a specific value in the canon of values. The range of risks of the individual options for action can be so wide that a differentiated preference for action is not possible, be it because of insufficient data, poor-quality data or the complexity of the overall system.

At the same time, the extension of standards to include ethical aspects should be considered for the certification of

21 “Risk-based” in English.

AI systems (for initial approaches see [45]). Standardization should not attempt to define a single set of values for all AI applications, since such a definition might not adequately reflect the existing relationships in the various fields of application.

#### 4.2.2.3.3 Privacy

As a special value, the principle of privacy protection is an expression of human dignity, autonomy and individual freedom. Accordingly, the protection of privacy through technology design and data protection law defaults (Art. 25 GDPR [95]) also has an ethical dimension, since value concepts and functional-cognitive aspects significantly shape the personality to be protected. Standardization should therefore promote technology design based on ethical criteria, both in the design and in the monitoring of AI applications, in order to protect the personal interests of users and those affected by the systems (in the sense of “privacy ethical design”).

So far, however, there is no uniform strategy in this regard. Standardization of AI applications is mainly limited to terminology and term definitions, the interoperability of AI systems, and security defaults. Standards which, in addition to ensuring interoperability and technical reliability, emphasize the consideration of ethical aspects and values in product or process design, or demand the responsible use of AI applications, are largely absent. Only in the area of medical and occupational safety law are these aspects also covered by product- and/or person-related organizational and documentation obligations. However, value-oriented requirements do not exist across all areas, although there would be opportunities to establish links: For example, within the framework of the ISO 9000 family of standards [105], [120] in the sense of an “explainable AI” it could be ensured that the interests of the “stakeholders” are sufficiently taken into account; likewise, when dealing with “risks” [120] the hitherto rather technical risks could be supplemented by ethical ones. In the future, standardization should close this gap, taking into account operationalizable values.

Standardization, which would be enriched by these ethical aspects in the sense of a “privacy ethical design”, would not only have the potential to become a European-wide benchmark by means of a collection of corresponding standards, but could at the same time also contribute significantly to increasing the acceptance of and trust in AI systems with regard to the integrated protection of privacy.

The principle of the protection of privacy through technology design and data protection law defaults (Art. 25 GDPR [95]) is an expression of human dignity, autonomy and individual freedom and therefore has an ethical dimension, since moral concepts and functional-cognitive aspects significantly shape the personality to be protected. Standardization should therefore promote technology design based on ethical criteria, both in the design and in the monitoring of AI applications, in order to protect the personal interests of users and those affected by the systems (in the sense of “privacy ethical design”).

#### 4.2.2.4 Criticality matrix of risk assessment

Due to its broad application, AI raises a multitude of legal and social questions. The use of AI-supported systems in HR, for example, makes it necessary to discuss labour and data protection issues from a new perspective. These efforts are to be supported. Since ADM systems can trigger far-reaching consequences through their decisions and their mistakes, such as discrimination (see [120]), it is important to make them sufficiently transparent and comprehensible.

As described in 4.1.2.1.4, learning AI systems achieve their essential functionality through the learning phase. As with humans, the testing of what has been learned is a great and new challenge for software development. Proof of this is not possible with learning AI systems today, since no method is known to completely verify/validate what has been learned – so there is still uncertainty as to whether the system meets all requirements and expectations in the operational environment. According to DIN EN ISO 9000 [105], the effects of uncertainty are called risks. DIN EN ISO 9001:2015 [120] emphasizes the risk-based<sup>22</sup> approach to thinking and acting for quality management systems (QM systems). In this context, it is necessary to continuously determine those factors that could cause their processes, products and services to deviate from the planned results and to implement preventive control measures. The systematic approach to a risk management system is presented in DIN ISO 31000 [93], in which, in accordance with ISO/IEC Guide 51 [121], risks are usually described in terms of the causes of the risk, the potential events, their effects and their probability. In order to avoid or limit the damage, which in the case of ethical criteria cannot be purely monetary, possible risk scenarios must be identi-

22 “Risk-based” in English.

fied and evaluated as early as the design and development phase. The evaluation must take place along the development process (see 4.2.2.1), always against previously defined criteria (goals), i.e. also against ethical criteria. Besides the lack of information, the interpretability of information can also play a role.

For safety-critical systems, the FMECA quality method is already used as a standard for non-learning systems. This method could also be applied to learning systems and related ethical criteria (AI FMECA) and could be used to identify factors that could cause unforeseen damage to the system as a whole that goes beyond a cost-benefit analysis, and to establish necessary transparency and traceability obligations regarding the decision logic. For this purpose, it is necessary to allocate the ADM system in its entirety to a criticality level within the scope of this risk assessment.

For the hazard prevention described above, it is essential to have clear criteria on the basis of which the respective ethical value or attribute can be made measurable. In a first approximation, this can be done with a comparatively rough classification, as mentioned above. An ordinal scale with values such as low, medium and high could be used. This is a procedure similar to that used in the corresponding standards for IT security (such as ISO/IEC 27001 [122]). The result of such a risk analysis must then be compared with the defined risk criteria. It must not be permissible to offset a low risk in relation to a normative value with a high or medium risk in relation to another and to arrive at a medium risk level overall. With regard to ethical values and the evaluation of algorithms according to ethical criteria, there can only be a maximum principle here, i.e. that a certain system or algorithm is no longer admissible if even one risk, with regard to an ethical principle, exceeds a certain limit.

For the effective application of these ethical standards for the purpose of creating and operating an automated decision-making system, it is necessary to make the corresponding ethical requirements operational. This requires the definition of criteria against which the degree of fulfilment (or of risk) of each individual criterion can be measured. Such criteria may come from the previously mentioned (examples of) cases of application, which have yet to be defined. There are current efforts to make ethical values in AI applications verifiable, and a methodology has recently been presented [123]. To ultimately operationalize the selected values, the goal must be to define the concept of “observable”, which can be used to check whether an ADM system meets a require-

ment. One possibility is the WKIO model (from the German **W**erte, **K**riterien, **I**ndikatoren, **O**bservablen = values, criteria, indicators, observables) which provides a systematic basis to concretize general values by breaking them down into criteria, indicators and finally measurable observables, thus making it possible to check whether an ADM system meets a requirement. Such a process of operationalization of values must also be accompanied by standardization in the coming years. As described in the introduction to this subchapter, in order to avoid discrimination and unpredictable long-term consequences, it is necessary, depending on the potential of the overall damage, to impose different transparency and traceability requirements on the decision-making logic of an ADM system.

Given the variety of ways in which ADM systems can be implemented and used, a differentiated regulatory approach appears necessary. IT systems with safety-related relevance follow ISO/IEC Guide 51 [121] and shall be designed, for example, according to ISO 12100 [124], [125], ISO 13849 [126], [127], ISO 14971 [128] or IEC 62061 [129]. For other IT systems the consideration can be done analogous to ISO 31000 [93] on the basis of a matrix-based risk assessment<sup>23</sup> by adapting regulatory provisions to different risks, e.g. in the financial sector (especially the Arrow II model, [131], [132]) or with regard to environmental risks [133].

The purpose of a risk matrix is not to identify concrete and hard thresholds between categories [133], since the conceptual distinction between risk classes cannot replace a thorough and detailed evaluation of concrete cases by a regulatory authority. Dealing with a particular source of risk is ultimately also a question of social values and risk tolerance, which entails a certain malleability and ambiguity. It is therefore hardly appropriate to draw rigid boundaries between risk categories. In addition, a practicable approach to AI ethics requires consideration of the respective application context, since this has an enormous influence on the potential risk, regardless of the technology used (cf. the use of a recommendation system in the context of online research and subsequent targeted advertising and recommendations for the choice of a suitable drug). Given the different applications and societal contexts in which ADM systems can be used, it is important that a solution can adapt to all needs when it comes to managing the risks of algorithmic systems [134]–[136].

23 Other governments are following the same approach, see [130].



The classification would have to take into account the total potential damage that an AI system can cause in its particular application context. Decisive factors in assessing this potential are the extent to which the ADM system may violate legal rights and human lives and limit the individual’s freedom of action. Here it can be seen that the use of a two-dimensional risk matrix, on which these factors describe the axes, simplifies the classification process without abstracting too much from the given complexity of an AI system [136].

#### 4.2.2.4.1 Activities of independent third parties

Third-party testing in the field of AI can be carried out in particular by conformity assessment bodies (CABs), which are outside of governmental bodies and not related to the users and manufacturers of the system. In the area of ISO standards, the CABs must define the necessary competence criteria for the personnel so that they are able to confirm conformity. If, in addition, legal framework conditions apply, these always take precedence (“law before standard” principle).

This can be shown in the example of the certification of a mechanical engineering enterprise as compared with the certification of a medical practice. The underlying standard is ISO 9001:2015 [120].

##### Case 1: Certification of a mechanical engineering company

On the basis of the ISO rules applicable to the CAB (in this case ISO/IEC 17021-1:2015, 7.1.2 [40]), the CAB is obliged to provide a mechanism for all persons involved in the certification process, which defines the competence criteria for the respective examiner (technical expert, auditor). In the present case, the CAB could therefore require that the examiner has the competence of an engineer or a comparable level of knowledge to examine a mechanical engineering company. Such regulations usually have an opening clause for people who have studied mechanical engineering but have not graduated, or for people who are not mechanical engineers but process engineers. If the CAB demonstrates this conclusively, persons who have not been trained to the qualification level of an engineer can also be called upon to audit such a company.

##### Case 2: Certification of a doctor’s surgery

Irrespective of how a CAB defines its own competence criteria here, German law stipulates that doctors are subject to professional secrecy (see professional code of conduct “MBO Ä”

1997 § 9 [137]) (comparable regulations also apply to tax consultants or lawyers). In this case, when examining the processes in a doctor’s surgery to determine whether they have been introduced in conformity with the standard, **only a doctor** or a person with an even higher level of training can be considered as a CAB examiner. All other persons, even if they are medically trained, cannot be granted insight into the processes of the practice by the doctor due to their professional secrecy. An examination can therefore not take place at all by law.

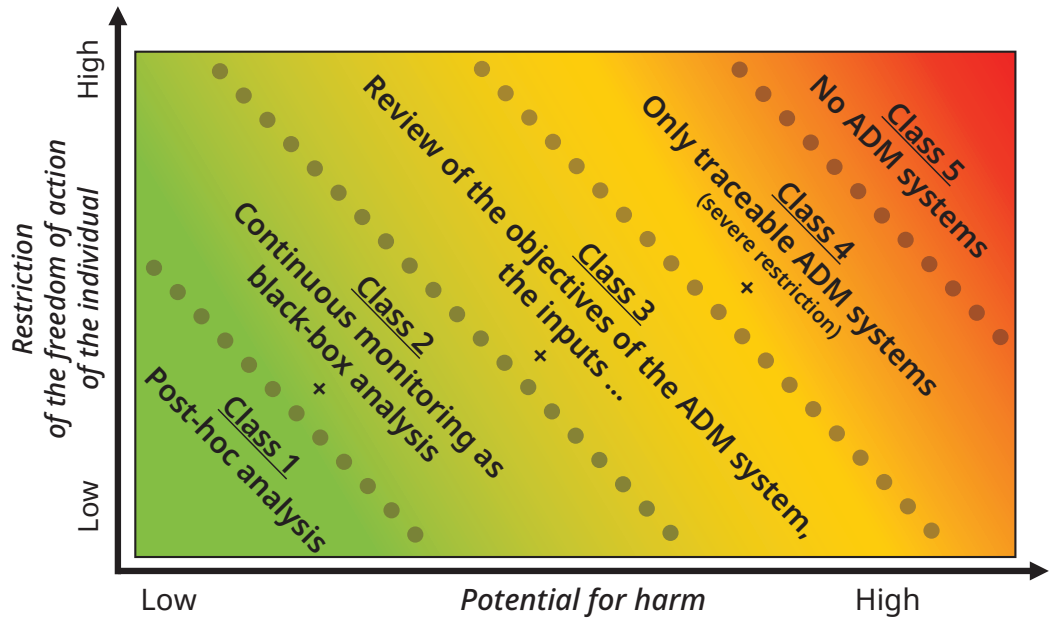
In this respect, it is important for regulation within the field of AI in key areas to have clear legal requirements or harmonized standards that ensure the **competence of the persons performing the tests**. Otherwise, an independent third party (a CAB) is allowed to define competence criteria in a way that may not be appropriate to the problem, even though it conforms to the standards.

It is essential to note that if an independent third party, i.e. a CAB, becomes active, it must be required that this takes place exclusively in the **accredited area**. Otherwise, there is no way of applying International Standards in such a way that the conditions described above can actually be required. This is only possible if the CAB itself is regularly audited by an independent organization, as is done in Germany by the German Accreditation Body (DAkkS) on the basis of the Accreditation Body Act. In this case a European-wide valid regulation would also exist, which would then take effect, since the corresponding accreditation body laws are equally valid and enforced in every EU country.

#### 4.2.2.4.2 Criticality model

A division into the two areas “high risk” and “no high risk” as demanded by the European Commission (EU Commission White Paper [15]) does not do justice to this complex problem. A differentiated approach will therefore prevail, which must be accompanied and shaped by standardization. The differentiated regulatory concept [136] developed by Krafft and Zweig, on the other hand, classifies ADM systems with their respective application context into one of five possible classes on the basis of two criteria and thus enables a rough assessment of the need for action (see Figure 17). A visual presentation of this concept was carried out by the Data Ethics Commission; details are given in 4.1.2.2.

**Figure 17:** Criticality model (as in [136])



**4.2.2.4.3** Examination of the necessity of a detailed criticality check

The economic viability of this concept is only given if the horizontal regulatory framework applicable to all ADM systems is kept to a minimum. In addition, sector-specific existing standards and guidelines in all (!) fields of application of artificial intelligence must be reviewed/revisioned/adapted and supplemented.

Current research indicates that only a small percentage of ADM systems currently in use in Germany have the potential to result in personal or social harm. In addition, there are studies that indicate that most of the damage only affects the private sector [138]. Nevertheless it is important to identify the systems with such potential. This can be done by means of a criticality check, which still has to be designed.

An ADM system is sorted by two axes into one of five categories. The higher the category, the more transparency and verifiability requirements are passed on to the decision-making logic of the system.

**Extent of possible violation of legal assets and human life (x-axis)**

For the x-axis, the critical aspect is the extent to which legal rights and human lives may be violated by an AI system. In order to assess this, at least the following must be considered:

→ **Impact on basic rights, equality or social justice:** Does an AI have a negative impact on the basic rights of a nat-

ural or legal person, are the mechanisms of social justice (e.g. pensions, health insurance) for a demographic group at risk or can the effects even be catastrophic and lead to loss of life (e.g. treatment of intensive care patients)?

→ **Number of persons affected:** Are a large number of persons affected (e.g. fair assessment in the case of a job application)?

→ **Impact on society:** Does the system bear the risk of affecting society as a whole (e.g. personalized selection of political news), regardless of directly perceivable damage?

In each case, it is impossible to assess the intensity of the potential damage by simply multiplying the amount of damage by the probability of occurrence. This would mean equating the risk of someone leaving the house without their umbrella in the event of an impending storm (high probability of occurrence, low damage potential) with the risk of a nuclear accident (low probability of occurrence, high damage potential). As the potential for harm increases, macro risks can arise that threaten our ability to act as a whole and are therefore unacceptable.

**Restriction of the individual’s freedom of action (y-axis)**

The y-axis shows the limitation of the freedom of action of the potentially affected individuals with respect to the algorithmic decision, and thus addresses the options for avoiding the potential for harm indicated on the x-axis. The better the chances of avoiding the possible negative consequences of a decision or the harm caused by it, the lower on the y-axis the ADM system would be. The three main factors that play a

role in assessing the dependence on the decision are control, selection and correction [123].

- Decisions and actions of an AI system that are additionally filtered by human interaction (e.g. the purchase of recommended items in an online store) imply a lower need for regulation than machines that act without human intermediaries (e.g. the emergency shutdown of a nuclear power plant). This aspect is summarized under control.
- The ability to exchange the AI system for another one (e.g. by changing a provider) or to avoid being exposed to an algorithmic decision at all is called **selection**. A one-sided dependency relationship between producers or operators and users, as well as monopolistic structures lead to dependence on one or a few systems. In the worst case, the user does not have the possibility to turn away from using certain services without being confronted with personal or social consequences (e.g. lack of access to health care, financial market).
- The importance of the possibility to challenge or have corrected an automatically generated decision, as well as the time needed for an adequate follow-up of the relevant application should not be underestimated. This is summarized by the term **correction**. Machine decisions that cannot be challenged at all increase the dependency on the decision. Repairing significant individual harm requires more time and effort than many cases with less harm. This aspect concerns the compensation for damage/liability, which is addressed in the dependence on the decision (y-axis).

#### 4.2.2.4.4 Risk classes

For systems that fall into Class 1, no transparency obligations would be required and no control processes would be permanently installed. In cases of doubt, a post-hoc analysis could be used to check for relevant damage. If the suspicion is confirmed, a new evaluation into a higher class would be conceivable.

In Class 2, the first transparency obligations would be required. To enable a “black-box analysis” [139], an appropriate interface must be provided for the system so that a controlling instance can check the input-output behaviour of the system. A description of how the system is embedded in the social decision-making process would also be necessary.

For systems in class 3, the input data should be described completely to a controlling instance. The stated quality (in the sense of numerical values describing the quality) of the decision system should be verifiable.

In class 4, all information about and decisions made by the software must be traceable and verifiable within a reasonable time, at least for a controlling instance. The demand for traceability generally excludes many learning processes (e.g. artificial neural networks), since they cannot fulfil this demand at the current state of research. All necessary interfaces would have to be provided.

Systems in class 5 should not be implemented. This class is justified by systems that are not compatible with the principles of democracy, such as evaluation systems based on continuous monitoring of the population, systems that override the presumption of innocence, or systems that have an approvingly lethal effect without human influence. Furthermore, systems that exceed a certain potential for harm and can only be implemented with a high error rate due to the difficult data situation (e.g. incomplete or faulty) would be in this class (e.g. identification systems for terrorists). This class does not exclude statistical methods that search for patterns in large amounts of data, but the finding of such patterns should not lead to unreflected decisions.

#### 4.2.3 Standardization needs

##### NEED 1:

##### Design initial criticality checks of AI systems quickly and easily

Unintended ethical problems and conflicts occur primarily in ADM systems with learning components that make decisions about people, their belongings or access to scarce resources, and have the potential to damage individual basic rights and/or basic democratic values. An initial criticality check as to whether a system can trigger such conflicts at all or whether it is an application far removed from any ethical issue must be made quick and easy by standardization. This horizontal, for all areas low-threshold check must quickly and legally clarify whether the system must meet transparency and traceability requirements at all. Especially with regard to the wide fields of application of artificial intelligence, such a risk-based criticality check in critical areas offers the possibility to make adequate demands and at the same time to counter the accusation of “ethical red taping” by developing completely uncritical fields of application free of additional requirements.

**NEED 2:****Operationalization of ethical values**

It is currently unclear how organizations that develop and use AI systems can measure and operationalize abstract ethical values. There are a number of promising approaches that have the potential to meet the challenge (such as the WKIO model), but the practical application of such approaches is still in its infancy. Open questions, problems and challenges can currently only be addressed to a limited extent, which is why standards offer the opportunity to transfer theoretical concepts for the operationalization of ethics into practice, to accompany them and to shape them consensually in dialogue with companies.

**NEED 3:****Standardization of a concept for privacy ethical design**

The principle of privacy protection is an expression of human dignity, autonomy and individual freedom, and an essential criterion for the acceptance of new systems. For this reason, standardization should promote the design of technology which safeguards the personal interests of users and affected parties in the sense of a “privacy ethical design”. This should take up and shape the previous approaches from the fields of medicine and occupational safety in a cross-divisional concept. This can be done within the framework of the project currently initiated in ISO/IEC JTC 1/SC 42 on an MSS for AI (4.1.3, Need 1 “Support for international standardization work on an MSS for AI”) by including the explainability of AI systems in the catalogue of requirements of the resulting document, and by extending the concept of risk to include ethical risks, as already done in the ISO/IEC 23894 Risk Management project.

**NEED 4:****Design of the value system**

Intelligent decisions based on general ethical principles require an examination of ethical values. If the machine knows the relation of meaning of different values and objects by means of an ontology, this is helpful. Autonomous systems must also be able to process unplanned situations. If, for example, the internal representation of objects is enriched by knowledge from an ontology when autonomous machines recognize the environment, this is a possibility to make a value system accessible to the machines. Ontologies allow the machine to create contexts without having to specify case patterns beforehand (as in W3C [140]). The research and subsequent standardization of the interfaces of ontologies to

consider ethical principles in concrete scenarios promises to meet the potential of the challenge.

**NEED 5:****Design earmarking of data**

Standardization should further shape the existing earmarking of data. This can ensure that there is documentation of the purpose for which the data was collected and can allow regulation of the conditions under which the data may or may not be used for other purposes.

**NEED 6:****Design interfaces for the AI development process**

The long development process of AI systems should be shaped by standardized interfaces. Here standardization can make an important contribution. These interfaces could include, for example, access to relevant training data sets and models of an AI system as a basis for external review. Primarily International Standards would promote the interchangeability of components and provide access for verifiers, and would ensure that requirements are met without much effort, thereby increasing confidence in the system.

**NEED 7:****Include quality backward chain in the AI life cycle**

It is recommended to include a quality backward chain with field data collection in the AI life cycle to identify and correct unethical behaviour during the application (see 4.3.2.3.2.4 Process checks: Quality assurance after delivery by product monitoring).

**NEED 8:****Design re-evaluation of AI systems**

AI systems should be widely used in a complex social context. A systematic process of ethical reflection and participation should therefore be initiated in AI development. Depending on the complexity of the AI and potential risks, several evaluation steps and a continuous involvement of interested parties, as well as ethics experts and ethically trained staff are recommended.

The background consists of a grid of hexagons. Some hexagons are light gray and contain a white outline of an award ribbon with a shield in the center. Other hexagons are dark gray. A central text box is overlaid on the grid.

## 4.3

# Quality, conformity assessment and certification

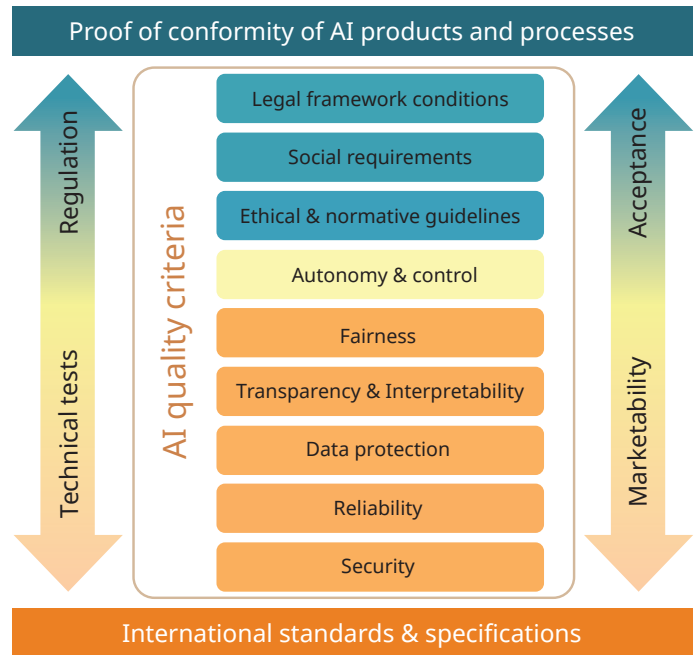
AI is increasingly being applied in different areas of everyday life (see also the chapters 4.5 to 4.7). Based on the assumption that AI can only unfold its full application potential if it is used according to high quality criteria, the following chapter deals with the resulting need for standardization with regard to quality criteria and their verification by a corresponding conformity assessment (based on the ISO/IEC 17000 series of standards [38]–[44]). A number of ideas discussed in this chapter can also be found in the *Impulspapier* and *White Paper* “Certification of AI Systems” issued by the Platform Learning Systems.

When testing AI systems, two levels can be distinguished (see Figure 18): On the one hand, assured properties of an AI system can be confirmed by a technical test. For example, the accuracy of a classification can be determined by precision and recall (technical level of testing). The second level is the evaluation level, which checks whether a system is suitable for a certain application (is the tested accuracy sufficient for the application?) or whether it meets certain ethical, legal or social requirements. A seal of approval [123] has been proposed for ethical considerations, which represents an interesting approach for the ethical evaluation of AI systems and is based on a value analysis procedure using a combination of target criteria, indicators and measurable variables. All tests of the second type should always be based on technical tests. It is to be expected that standards and specifications can be formulated primarily for the first level of testing, but that questions of the second level of testing are often the subject of regulation or social discourse.

Such conformity assessments can be carried out by the manufacturer itself, the buyer or an accredited third party body. In the course of conformity assessment, products, systems and processes may be subject to testing, calibration, validation, verification and certification or inspection. In certain areas (such as in accordance with the EU Medical Devices Regulation [141]), certification by a Notified Body is even mandatory prior to placing the product on the market.

Certification is carried out within the framework of conformity assessment by a third party according to the applied conformity assessment programme.

According to international expert commissions, such as AI HLEG, proof of conformity for AI products and processes is based on the following normative, legal and technical quality criteria (cf. also IAIS White Paper) [45].



**Figure 18:** Classification of the categories of AI quality criteria in conformity assessment [45]

**Law, society and ethics**

AI applications have a disruptive potential. Conformity with social, ethical and legal frameworks mainly serves the protection of legal or ethical fundamental interests of persons (4.2.2.3). The AI conformity checks in these categories are intended to prevent and help to avoid impairments of groups and individuals, injustice or ethically unjustified conditions of society.

**Autonomy and control**

AI applications increasingly work autonomously, i.e. they pursue a given goal while freely choosing the means to achieve it. The AI system is free to choose the means, but not the actual objective. In this context one speaks misleadingly of the “autonomy of action” of the system, although the objective is not changed. This analogy gives rise to an area of conflict regarding the autonomy of humans, since such AI applications can in turn influence humans in their choice of goals and means. AI conformity tests must be able to make statements about autonomy and control at the interface to the technical AI system if the AI application interacts with human decision-making, for example, by generating decision proposals, generating and possibly executing control commands, communicating with humans or being integrated into work processes.

The following quality categories are part of the technical testing of AI systems.

### Fairness and non-discrimination

AI applications learn from historical data, which is not necessarily unprejudiced. In order to avoid unjustified unequal treatment in an AI application and to exclude undue discrimination, AI applications must be verifiable to ensure that individuals are not discriminated against in social outcomes because they belong to a marginalized or discriminated group (see 4.2.2.2).

### Transparency and interpretability

The transparency of an AI application can significantly contribute to its acceptance. For this purpose, information on the correct use of the AI application must be available. Essentially, requirements for interpretability, traceability and reproducibility of results must be checked, requiring insights into the inner processes of the AI application. There is still a considerable need for research into the colloquially associated demand for the explainability of an AI application, even if the explainability of the effects of AI-specific technological features is limited.

### Data protection

The technical examination of the data protection regulations, in particular the GDPR [95], the BDSG [142] and the requirements of the Hambach Declaration [143], must be observed for AI conformity tests.

### Reliability

From a technical point of view, testing the reliability of an AI system includes requirements for correctness, traceability, assessment of the uncertainty of results, and of the robustness against attacks, errors, and unexpected situations and thus overlaps with the concept of security in the narrower sense. Tests of the reliability and security of AI applications are essential basic requirements to make statements about their trustworthiness.

### Security/safety

The security of AI applications includes security against threats and attacks and functional safety in the broadest sense. The security/safety of AI systems is discussed in detail in Chapter 4.4. Reliability, data protection and data security are also taken into account. In terms of testing methods, it should be noted that the technical test bases for AI systems must be developed and related to existing test procedures.

## 4.3.1 Status quo

In the following the essential terms of objects and activities of conformity assessment are listed.

### 4.3.1.1 Conformity assessment

Demonstration that specified requirements are met (ISO/IEC 17000 [38]). Defined requirements (i.e. needs or expectations) can be detailed (e.g. concrete technical specifications) or general (e.g. safe, robust, transparent, fair).

To differentiate the objects of conformity assessment:

1. Product (e.g. hardware, software)
2. Process
3. System
4. Service
5. Management system:
6. Person
7. Information (e.g. declarations, assertions, predictions)

Objects of a conformity assessment can also be combinations of these individual objects (e.g. development process + product, product + service, system + assertion). The specified requirements must be clearly assigned to the object (e.g. technical specification for the hardware, fairness criteria for the process, robustness of a system, competence requirements for a person, plausibility conditions for an assertion).

To differentiate the activities:

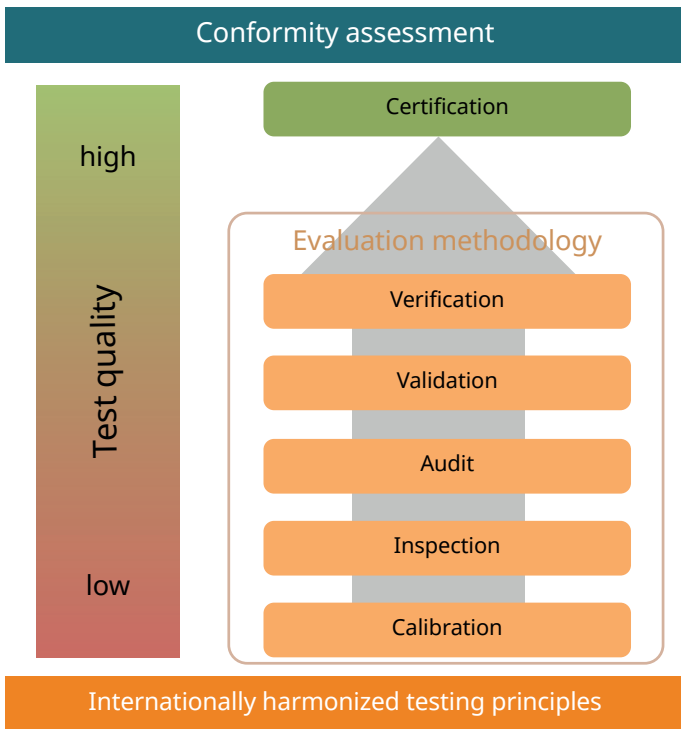
By clearly assigning the specified requirements to defined objects (see above), the activities for “selection” and “determination” (see process of conformity assessment) can be determined. Their results may be sufficient for the given situation (e.g. for analysis or characterization) or may subsequently be subject to “assessment” with a view to a “decision” on conformity of the object.

#### 4.3.1.1.1 Types of conformity assessment

In the following, the types of conformity assessment (see Figure 19) are described.

### Testing

Determination of one or more characteristics of an object of conformity assessment by a procedure. The procedure may be intended to control variables within the test as a contribution to the accuracy or reliability of the results. The results of the test can be presented in the form of specified units or objective comparisons with agreed references. The result of the test may include comments (e.g. opinions and interpretations) on the test results and compliance with the specified requirements.



**Figure 19:** Evaluation methodology and test quality

### Calibration

Activity which, under specified conditions, in a first step establishes a relationship between the quantity values provided by standards with their measurement uncertainties and the corresponding displays with their associated measurement uncertainties, and in a second step uses this information to establish a relationship with the aid of which a measurement result is obtained from a display.

The result of a calibration can be expressed in the form of a specification, a calibration function, a calibration diagram, a calibration curve or a calibration table. In some cases it can consist of an additive or multiplicative correction of the display with the assigned measurement uncertainty. Calibra-

tion should not be confused with adjustment of a measuring system, which is often wrongly called “self-calibration”, nor with verification of the calibration. Often only the first step in this definition is considered as calibration [144].

### 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. An examination may include direct or indirect observations, which may involve measurements or reading of measuring instruments. Inspections can be limited to examinations in conformity assessment programs or contracts.

### Audit

Check that an organization’s processes, practices and procedures meet certain requirements formulated in a standard (e.g. an MSS, see 4.1.2.2.3). This check is usually based on a list of criteria derived from the underlying standard, which describes how requirements are checked. Audits include the inspection of documentation provided by the organization to be audited, interviews by the auditor, but also on-site inspections.

ISO differentiates between three levels of the audit:

- Audit by the organization to which the audit refers (self-disclosure);
- Audit by a customer, supplier or partner of the organization to be audited;
- Audit by an independent third party. Such an audit can lead to certification.

ISO 19011 [145] provides guidelines for audit planning, audit execution and audit follow-up.

### Validation

Confirmation of the plausibility of a specific use or application purpose by providing objective evidence that specified requirements have been met. Validation can be applied to assertions to confirm the information provided by an assertion in relation to its intended future use.

### Verification

Confirmation of truthfulness by providing objective evidence that specified requirements have been met. Verification can



be applied to assertions in order to confirm the information provided by an assertion that relates to events that have already occurred or that relates to results that have already been obtained.

## Certification

Confirmation by a third party relating to an object of conformity assessment (accreditation excluded). A “third party” is independent of the supplier of the object of the conformity assessment activity and has no interest as a user. Testing, inspection and validation/verification activities may also be performed by the supplier (first party) of the object to be evaluated or by a person/organization with an interest as a user of that object (second party). Certifications are only offered by independent bodies.

### 4.3.1.1.2 Conformity assessment process

Conformity assessment is divided into five phases:

- **Selection** = Selection of applicable requirements, choice of methods, planning, sampling
- **Determination** = Activities to collect evidence of conformity with regard to the specified requirements, i.e. analyses, tests, evaluations, investigations, audits, tests, inspections, validations, verifications, etc.
- **Review** = Conclusion regarding suitability, adequacy and the sufficient amount of evidence collected
- **Decision** = Deciding whether or not the assessed object has been shown to conform to the specified requirements
- **Attestation** = Formal issue of the statement of conformity, e.g. test report (test passed/failed) or certificates

### 4.3.1.1.3 Types of conformity assessment bodies:

Depending on the type of conformity assessment, the ISO/IEC 17000 series distinguishes between different types of assessment bodies which, according to the activities listed above, inspect, analyze, test or measure product safety and quality and objects of protection:

- **Testing laboratory** (ISO/IEC 17025 [42])
- **Inspection body** (ISO/IEC 17020 [39])
- **Validation/Verification body** (ISO/IEC 17029 [43])
- **Certification body** (ISO/IEC 17021-1 [40] for management systems, ISO/IEC 17024 [41] for persons and ISO/IEC 17065 [44] for products, processes and services)

### 4.3.1.2 Existing standards and specifications from other areas with relevance for AI quality and conformity assessment

AI applications are usually implemented as components of larger IT systems. These AI components can be realized by a variety of different technologies. These AI applications are used in many industrial and everyday applications, where the actual AI component often interacts with other software, information technology, mechanical and electronic modules of the overall system.

The first step in standardization is therefore to identify existing standards and specifications that are relevant to the quality (and checking) of these systems. Standards from the areas of software (AI component), IT security (overall IT system), data quality and functional safety (application context) are particularly worthy of consideration.

Table 13 in Chapter 6.4 shows national and global standardization committees. Working groups relevant for quality, conformity assessment and certification are marked in the column “Relevance for quality, conformity assessment and certification (4.3)”.

In principle, any standard that formulates requirements for a software application is also relevant for AI components as a special software component, regardless of the technology used. It must first be checked which standards already sufficiently cover the AI-specific properties and whether additions or changes are necessary.

Some prominent examples from the fields of software development and functional safety are listed below. However, this list makes no claim to completeness. In addition, there are very many relevant standards on IT security which are discussed in detail in Chapter 4.4. In addition, standards focusing on AI are being revised in this and other areas to address AI-relevant aspects. Table 11 in Chapter 6.2 gives an overview of standards and specifications in different thematic areas that do not yet provide detailed information on the application of AI components. The standards that formulate relevant requirements and quality criteria for software are marked in the column “Relevance for quality, conformity assessment and certification (4.3)”.

#### 4.3.1.2.1 Software development

AI processes can be integrated into existing software development standards such as ISO/IEC/IEEE 12207 [58] (Software life cycle processes), ISO/IEC 27034 [71]–[78] (Application Security) and ISO/IEC 25010 [146] (System and software quality models). For example, a test of trained AI-based software systems for “functional safety, efficiency, transferability, maintainability and reliability” can be carried out according to ISO/IEC 25010 [147].

#### 4.3.1.2.2 Functional safety

The IEC 61508 [79]–[86] series of standards defines requirements for the various life cycle phases of electrical, electronic and programmable electronic (E/E/PE) systems that perform safety-related functions. IEC 61508-3 places a special focus on the requirements for the development of safety-relevant software. This also includes requirements for the tools used in the development process. Four safety integrity levels (SIL) are defined as a measure for the necessary risk-reducing effectiveness of safety functions and the resulting requirements on the safety-relevant system. Until now, the use of AI functionality is not recommended but also not excluded by IEC 61508. The responsible committee IEC/SC 65A, however, is considering the topic of AI for an update of IEC 61508 and is working together with ISO/IEC JTC1 SC42. There, a technical report on functional safety and AI systems is under development. IEC 61508 has a broad acceptance and application in industry and is the basis for several application-specific standards, e.g. for the process industry, mechanical engineering, control technology in nuclear power plants and railway signalling technology. The specification ISO/PAS 21448 [148] describes the safety of the target function and also includes performance restrictions that have their origin in environmental influences or communication. The standard ISO 12100 [124], [125] defines general principles and methods of machine safety as a basic safety standard, but is not a functional safety standard in the narrower sense.

#### 4.3.1.2.3 Data quality

Since the quality of an AI component is closely linked to data quality, standards on data quality and big data are also listed in Table 11. DIN ISO/IEC 25012 [88] introduces a model of data quality. ISO/IEC 20546 [34] and ISO/IEC TR 20547-2 [149] and -5 [150] deal with big data, its terminology and reference architectures.

#### 4.3.1.3 Existing standards and specifications on AI quality and conformity assessment

Table 10 in Chapter 6.1 lists existing standards and specifications that deal explicitly with AI applications. The standards that formulate relevant requirements are marked in the column “Relevance for quality, conformity assessment and certification (4.3)”. This list is not exhaustive, but from today’s perspective it represents the majority of the relevant standards and specifications.

In Germany, DIN has published two DIN SPECS in which a quality meta model for AI (DIN SPEC 92001 [87]) and a guide for deep learning image recognition systems (DIN SPEC 13266 [151]) are presented. At European level, ETSI addresses artificial intelligence in technical specifications relating to emotion recognition (ETSI TS 103 296 [152]) and autonomous networks (ETSI TS 103 195-2 [153]). At international level, the ITU-T focuses, within the published standards on requirements (Y.3170 [154]) and AI capabilities (Y.3173 [155]) with regard to AI in future networks. Within published documents, the consortia IEEE and UL deal with the assessment of autonomous systems (IEEE 7010-2020 [156] and UL 4600 [157]).

Apart from DIN SPEC 92001-1 [87] all specifications mentioned and listed in the table deal with AI components related to a concrete application. Work on a number of standards dealing with the quality of AI systems is in progress at international level, for example in ISO/IEC JTC 1/SC 42. In addition, IEEE standards are also in preparation or available, as are DIN SPEC 92001-1 and SPEC 92001-2. In further standardization activities, the quality criteria mentioned there would have to be compared with the quality criteria of ISO/IEC 25010 [146] and the AI-specific requirements would have to be highlighted.

#### 4.3.1.4 Standardization activities with relevance for AI quality and conformity assessment

Table 12 in Chapter 6.3 lists standardization activities relevant to AI quality and conformity assessment in the column “Relevance for quality, conformity assessment and certification (4.3)”. This list is not exhaustive, but from today’s perspective it represents the majority of the relevant standardization projects.

Currently, numerous activities for AI standardization are taking place on all levels of standardization. Especially the work in ISO/IEC NP 5059 is to be emphasized, because here work is done on quality requirements for AI following the software quality requirements of ISO/IEC 25010 [146].

## 4.3.2 Requirements, challenges

### 4.3.2.1 Need for testing and marketability

In April 2019 the AI HLEG published ethical, legal and technical key requirements for trustworthy AI-based systems [22]. In most cases, these are hybrid applications, i.e. they consist of AI components and non AI-based software and hardware, and are basically understood as special IT. The user industry in Europe expects the market-driven development of criteria and methods for the technical testing of I systems. There follows a discussion of the scope of such testing.

#### 4.3.2.2 Scope of a test

In this chapter we discuss which aspects should be considered in the context of a test of an AI system. This includes the components of an AI system, as well as AI-specific challenges that arise when testing these systems.

Further quality requirements result from the fact that AI systems are often also components of a larger product (e.g. the Platform Economy) for whose interoperability standards must also be set to ensure additional connectivity and interchangeability in the end product. Without such guarantees, global interaction is almost impossible, and this ultimately also prevents the scalability of solutions.

With regard to the aspect of the examination of quality criteria there is partly a great affinity to test procedures of functional safety, software development and IT security, which can be attributed to the fact that AI applications are hybrid IT systems. The marketability of a potential test method that addresses the above-mentioned aspects therefore requires an integrated approach that extends existing test methods to AI-specific criteria. There follows an analysis of the components of an AI system that require consideration when testing. In addition, the AI-specific challenges that need to be addressed to close the gap illustrated above are described in the following.

#### 4.3.2.2.1 Components of an AI system

Components of an AI system include algorithms, databases and interfaces to the overall system. In principle, AI-based system components are based on symbolic and sub-symbolic methods of artificial intelligence. These include techniques for decision-making (e.g. decision-theoretical expert systems), knowledge representation (e.g. ontologies and knowledge graphs), methods for applying knowledge (e.g. logical reasoning and probabilistic methods) and machine learning methods (e.g. supervised learning and unsupervised learning). A detailed description of the classification of AI components can be found in 4.1.2.

Here the methods of artificial intelligence in an AI application can be realized by software. Depending on the capability spectrum of an AI application, hybrid methods (e.g. hybrid neural network models) can also be used in which symbolic and sub-symbolic techniques are combined. In AI applications there is an adaptability (dynamics) of the partial components of methods of artificial intelligence. For example, in machine learning processes, activation, transfer and summation functions determine the dynamics of a neural network [158]. On the other hand, a dynamic can manifest itself in the changeability of knowledge through AI methods, for example based on the AGM theory [159], [160].

Regardless of the actual realization of the AI application, a quality assessment should include the following aspects:

→ The quality of the data used: This includes, among other things, a possible bias in the data, which can negatively affect the fairness of the overall system, and the integrity of the data, since these significantly determine the behaviour of an AI component and thus make it necessary to secure the training data sets against indirect attacks through their manipulation. This applies in particular to continuously learning (self-learning) systems that are further trained in the field and whose input data for continuous learning is not under the direct control of the manufacturer. Therefore, quality assurance of the data supply chain itself is also necessary, as it plays an essential role with regard to the quality aspects of the data. Also, the data used for the training of a model and its distribution (e.g. image resolution, statistical distribution) must correspond to the operational environment.

Synthetically generated data are increasingly being used in the development of AI systems. Here, an artificial representation of an original data set is created, which has the most important statistical properties of the original

data set. Such synthetically generated data sets are especially helpful if either the amount of original data is too small (an example is the training of ML models for autonomous driving) or if the original data contains sensitive personal characteristics. The quality of such synthetically generated data is measurable and should meet the same quality requirements as real data sets.

- The selection of the method/algorithms, their hyperparameters and the evaluation of a learned model. In general, empirical methods of testing as well as verification methods for quality assurance of a trained system are suitable here. Both are subject to the challenges described in the following chapter. To check the quality, it is necessary to consider alternative hyperparameters and their influence on the quality. For parameter selection and other areas of engineering, methods of automated machine learning under the term AutoML are in development and first use, among others as a service.
- Also, the assessment of the overall IT system in which the AI component is embedded. This results in particular in interfaces to other technical IT environments such as cloud architectures, server farms, data repositories and data supply chains, statistical analysis packages, etc.
- The man-versus-machine interface. Here, human and machine factors need to be considered. But machine-versus-machine and AI-system-versus-AI-system require validation. The interface can be facilitated by the AI system giving the human feedback explaining what it has “understood”.
- The behaviour of the AI application after delivery during its use in the operational environment (product observation), until the end of its life cycle (see 4.3.2.3.2.1 and 4.3.2.3.2.4)

#### 4.3.2.2.2 AI-specific challenges

In contrast to conventional IT systems, AI applications have some special features for which quality criteria and test methods must be established and which pose substantial challenges for existing and future test methods. This includes:

- A correctness term for KI systems: Rule-based algorithms have a clear source code that can be tested using classical methods. Examples of suitable verification methods are classical proving and proof assistants. Certain definite parameters can also be appropriately tested. With learning systems, not only the software architecture (e.g. NN model selection) and the source code quality are involved, but

also what has been learned. Also, in contrast to classical systems, AI-based systems often work statistically and will therefore not achieve an accuracy of 100% of the specified behaviour. Therefore, a test of AI-based systems must define a sufficient requirement for accuracy and aim to argue for this requirement on the one hand and for the remaining cases to secure the system by further measures, e.g. safeguards. It remains that certain residual risks can be tolerated as part of the application-specific risk management.

- Dynamics of AI systems: AI systems, which are based on machine learning methods, are often subject to a dynamic during operation that has two causes: On the one hand, the operating environment can change so that the originally learned model only inadequately reflects reality (concept drift). On the other hand, the model can continue learning during operation, for example through user feedback. This is designated model drift. For a potential AI test, this means that the result about the assured properties of the AI system need not be valid at a later date. This represents another central difference to the verification of conventional software. The following measures are conceivable to counter this problem: 1) Model drift can be avoided by introducing structured model updates. Potential quality requirements can be defined for such updates, so that the assured and tested properties are maintained after the update. 2) Similar to cloud certifications, a continuous test of the AI system by monitoring suitable KPIs is conceivable. However, a suitable selection of such KPIs is currently still the subject of research and development. Alternatively, suitable measures, e.g. safeguards, can be taken to ensure that the system cannot assume critical states. 3) Possible uncontrollable behaviour of AI systems can be prevented by involving a supervisor (human-in-the loop).
- Uncertainty: Uncertainty regarding the correctness of an output is an intrinsic property of data-driven AI applications. Apart from the simple observation that the application of a model created by a machine learning process to a new, so far unknown input can lead to a correct or even incorrect result, research on the uncertainty of models in the narrower sense is concerned with the view that a learned model can be regarded as a probabilistic function, and thus each statement made by the model is provided in principle with a confidence, the knowledge of which in turn allows various conclusions regarding the use of the model in a given case. Unfortunately, for com-

plex learned models, the actual valid confidence values are not directly visible. The situation is complicated by the fact that the uncertainties arising in the application of the model can be caused not only by different aspects, but also by interacting aspects: insufficient or imprecise data, limitations in the expressiveness of the chosen model class, or an immanent, non-deterministic behaviour of the modelled objective function (e.g. long-term weather forecast). Accordingly, a precise knowledge of the model uncertainty would in turn allow conclusions to be drawn about the data situation, model complexity and prediction quality in the application. The latter in turn is a central element of layered security architectures, where alternative mechanisms are applied at upper levels (e.g. driver takes the wheel), if the AI application on the lower level signals too much uncertainty (monitoring approach). There is a broad spectrum of research approaches to capture the uncertainties associated with a learned model under restrictive conditions, ranging from simple subsequent “model calibration” and targeted interventions in the actual learning process to complex redundancy procedures and more or less holistic mathematical analyses. In view of the ever-increasing model complexity and breadth of applications of learned models in safety-critical areas as well, the development of efficient, precisely effective and generally applicable methods for determining and testing the uncertainty of models is urgently required.

- Transparency/traceability: An AI system is transparent if its genesis and mode of action are presented openly, completely and understandably. This includes particularly the data basis and the algorithmic component. The decisions/proposals of an AI system are traceable if the factors that led to their creation can be understood by a person. The following aspects in particular play an important role in transparency: Transparency of the data used for training, the annotation of the data (e.g. inter-annotator agreement using Cohen’s kappa or Fleiss’ kappa). Transparency in the selection of methods. Transparency and traceability of results (influence weighting of the entered variables). Transparency in the approach (e.g. through a history of the hypotheses tested during parameter optimization or model generation). Transparency in the secured application (i.e. when a model can make sound decisions or when it operates outside or in peripheral areas of the input data). In general, a distinction must be made between transparency for the end user and interpretability.

From a technical point of view the question of basic transparency is not easy to answer, and the tension between higher accuracy or robustness and the explainability of models is a well-known dilemma in the AI world. Although “black box” models are in many cases more accurate or more robust than, for example, rule-based models, they are only conditionally interpretable. In part, this explainability can also be achieved by downstream procedures, such as training of explanatory models or an analysis of the input/output behaviour of models, so-called Local Interpretable Model-agnostic Explanations (LIME) analysis. Currently, the interpretability of models is an active field of research and many efforts are being made to better understand the learning processes of “black box” models, to visualize their internal processes and to explain the resulting decisions.

- IT security: AI components and AI-based systems are now exposed to IT security risks such as adversarial attacks. Since these often work statistically and their mode of operation is not yet fully understood, quality assurance poses major problems for the IT security of AI components. Modifications of data that are imperceptible to humans, e.g. in images, lead to misclassifications when using adversarial samples, e.g. by subtle manipulation of traffic signs on the road or by adding targeted noise in already existing images. AI systems themselves and the models they contain are also subject to IT security risks. The trained model represents a business value to be protected and must therefore be protected against reverse engineering and its training data. Corresponding attacks can also have an impact on data protection, since techniques already exist that allow the extraction of individual training data records. Detailed explanations can be found in [Chapter 4.4](#).
- Hyperparameters: In addition to the selected AI method or algorithm and the data used for training and testing, the associated hyperparameters significantly determine its quality and can lead to effects such as overfitting, where the system achieves a particularly high level of accuracy for the training data, but only a low level of accuracy in operation. Hyperparameters include properties of the model regarding its size (e.g. number of layers of a deep neural network) as well as learning parameters like the number of epochs and the learning rate.

### 4.3.2.3 Test methods

#### 4.3.2.3.1 Verification of AI systems

Suitable verification is the basis of any conformity assessment in the development of systems. The specific challenges of learning systems mentioned in the Introduction also make demands on the verification during development. Beside a corresponding documentation with configuration management, here the consideration lies on testing (check against criteria), verification (formal check of the AI module against the specification) and validation (formal check of the application in the use environment).

The “Product Quality Model” of ISO/IEC 25010 [146], addresses two fields of topics:

1. Functional testing: “what the system does”
2. Non-functional testing: “how the system does it”.

For a test environment, an application-specific and boundary condition-dependent action framework should be created, within which test methods can be defined. The test procedures and test depths depend on the identification of relevant user groups such as developers and users, as well as application scenarios, data protection and potential for harm (see Figure 8). The boundary conditions, structures and interfaces of a possible location of the AI-based system should be simulated within the test environment. The test environment should be separated from the external environment so that distortions in results can be avoided. In the test environment, sequences for tests of different depths should be guaranteed. The test depth can be determined based on deployment risk, complexity of the AI application, effort and cost. In order to demonstrate the conformity of a system in a traceable way (conformity assessment), it is necessary to define the underlying requirements unambiguously. For systems based on AI, a catalogue of requirements should be developed in which aspects such as system requirements, system architecture, software requirements, software architecture, source code structure, module structure, software integration, software quality, training and test data quality, system integration and system quality are documented [161]. On the basis of a framework for action, AI methods and capabilities can be subjected to a conformity assessment with a view to appropriate suitability depending on quality and validation requirements, taking into account ethical, legal and social assessment schemes. AI applications can be described based on the criticality pyramid in 4.1.2. Similar to the determination of measurement capabilities in the calibration/testing of

measuring instruments using traceable, validated metrological standards and references, reference data, benchmarks and reference methods can be an important part of the test in certain areas. For example, benchmarks validated in ECG analysis can be performed with test data not previously known with the AI method and compared with the results of reference methods.

When testing AI systems, two approaches can be followed: Process tests can be used to verify quality standards for the operation and development of the AI system, while product tests verify assured properties of AI systems. Both test approaches must fit into an overarching testing framework that ensures the comparability of tests of different AI systems. This testing framework should be open with regard to the selection of subsequent test methods, but should be connectable and compatible with established test methods. Examples of established test methods are the conformity assessment of the New Legislative Framework (NLF) or CC [47].

#### 4.3.2.3.2 Process tests

The product AI brings a host of new challenges. Among other things, depending on the method used, transparency or traceability is limited with regard to a decision made by an AI. Therefore, transparency with regard to the AI development process in the form of a process test is even more important [162]. It should also be noted that AI products are often provided in the form of Internet-based services or access Internet and cloud services. Such services are often continuously updated: A test of AI products, especially service-based AI products, should therefore be supplemented by an audit of the processes of the organization providing these products. In addition, organizations using AI products should also be able to obtain proof of their responsible use of such technologies, for example in the form of a test report or an appropriate certificate.

#### 4.3.2.3.2.1 Assessment of the consequences of using AI

At the beginning of these processes, in addition to the requirements adapted to the AI challenges, there is also a consideration of the expectations, demands and fears of other affected parties, e.g. customers and partners of an organization, end users of AI products, etc. Organizations should be able to understand the impact and consequences

of using such products and, if necessary, to reconcile them with their own objectives: Such an extended management of risks of the use of AI, which considers not only risks for an organization but also the impact on third parties, should be implemented and verifiably documented by appropriate management functions, roles and responsibilities.

#### 4.3.2.3.2.2 Development process of AI systems

Transparency with regard to the AI development process in the form of a process test should include the documentation of important decisions regarding the selection of certain criteria and indicators (e.g. metrics, accuracy, precision, recall, specificity and sensitivity). Furthermore, requirements for continuously learning AI systems should be appropriately designed (goal alignment) and documented. Since the training process has a significant influence on the quality of an AI, the training progress must be ensured. This requires a versioning of the software, including the data used for training. In addition to the versioning of the software, the documentation of central and system-relevant decisions is important, e.g. decisions and decision changes regarding model selection, data preparation (feature engineering) and the classification into training and test data. Before validating an AI, the documentation for testing and verification must be completed.

#### 4.3.2.3.2.3 Use of AI systems and their provision as services

Processes involved in the use of AI systems, particularly in their provision as services, include the continuous review and evaluation of performance and security metrics, the determination of appropriate responses to incidents, and the establishment of appropriate countermeasures. In addition to these generic processes, AI must also be considered and supported by appropriate management processes, e.g:

- The impact of automated decisions made by AI systems and the resulting loss of control.
- The loss of organizational knowledge that can be caused by the use of automated decision-making systems and the resulting strong attachment to such systems (“blind trust”).
- The possibility that services of third parties are used for purposes that are questionable within the ethical self-image of an organization.
- Dealing with limited transparency and explainability of AI systems.

Process tests should be based on established MSS (e.g. ISO 9001 [120], ISO/IEC 27001 [122], ISO/IEC 27701 [163], etc.); however, such standards, as far as they are currently published, only cover parts of the development and use of AI (quality, security, data protection, etc.) The development of a stand-alone MSS for AI, which has already been discussed in Chapter 4.1, is therefore recommended. This can be used in addition to product testing (see 4.3.2.3.3) for conformity assessment and certification as a result of an audit.

#### 4.3.2.3.2.4 Quality assurance after delivery through product monitoring

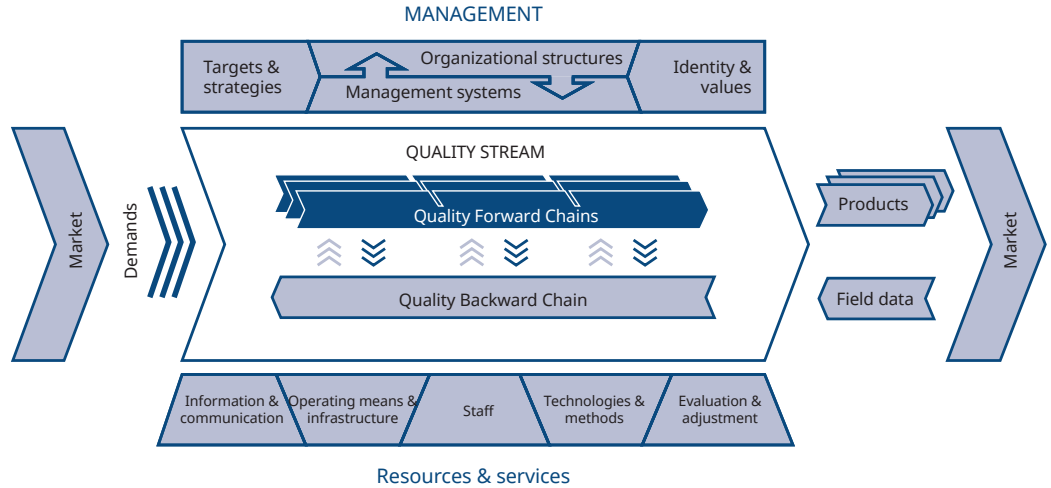
As a quality assurance measure during operational use in the AI life cycle (see 4.1.2.3.1), an active product observation with evaluation of acquired field data should be normatively defined for AI systems, as is already practised for safety-critical systems in the automotive [164], aerospace and defence [165] industries.

To ensure that recognized problems and risks in the application of an AI system, e.g. unethical or unnecessarily endangering behaviour in the operational environment, lead to appropriate corrective and sustainable improvement measures, it is necessary to feed back the quality-relevant information gained through product monitoring to the corresponding point of action in the AI life cycle. This feedback must include the possibility of warnings to customers and authorities, as well as a product recall.

A systematic feedback for product, system and process improvement through internal [166] and external [167] quality-relevant information is described, for example, in the “Aachen QM Model (AQMM)” [168] as a “Quality Backward Chain” (see Figure 20).

For AI systems, this product observation can be done by storing evaluation results together with the corresponding source data (e.g. sensor data), transferring them to the AI development company and evaluating them there. In commercial aircraft today, this is done partly during flight or when the aircraft is electrically coupled to the gate at the airport, with around 1 GB of data per flight hour. Current motor vehicles also have integrated mobile radio interfaces for transmitting rudimentary field data to the vehicle manufacturer and, in some cases, digital data loggers in order to record at least a certain period of time before a damage event. For AI systems, the amount of data required can be very large, so a workable

**Figure 20:** The Aachen QM Model with “field data” and “Quality Backward Chain” [168]



solution for each AI application must be determined during development.

**4.3.2.3.3 Product tests**

In addition to process tests, which ensure compliance with good standards for the development and operation of AI applications, there is a need for product tests which test the properties of an AI system itself. On the one hand, the test can include the product properties assured by the developer, on the other hand it can confirm compliance with certain industry- or product- specific standards.

For such a product test, a framework is required according to which the functionalities of the AI application can be specified uniformly. In addition, evaluation principles are required which indicate when the promised functionalities can be considered fulfilled. These evaluation principles should particularly include an overview of common metrics that make performance measurable with respect to different technical properties (e.g. robustness against adversarial attacks, reliability etc.). The challenge here is that the adequacy of the metrics used can be highly dependent on the usage context or use case. Some AI models (e.g. word embeddings) do not have their own quality criterion, but can only be compared in the application by further procedures. A solution should be found for this as well.

**Example of a use case from the mobility sector autonomous driving**

These issues are currently being investigated in initial pilot projects, such as the Federal Ministry for Economic Affairs and Energy (BMWi) project AI security, and are expected to yield valuable findings for the standardization of AI systems:

The goal of the AI security project is the development and investigation of methods and measures for the security of AI-based driving functions for the use case “pedestrian detection”. The knowledge gained should make it possible to better determine and assess the technology. In addition, this is intended to create a stringent chain of argumentation which, from the expert’s point of view, justifies the security of AI functions. Ultimately, communication with normative committees and certification bodies should support an industry consensus on an AI testing strategy.

Furthermore, it requires the specification of different levels of trustworthiness (see the criticality pyramid in 4.1.2.1.5 or the risk criticality model in 4.4.1.2), which are confirmed by an audit according to scope and depth. For this purpose it is necessary to define a suitable framework at different test depths. The range of methods includes document checks, audits, black box and white box tests, as well as validation and verification.

To carry out such product tests, suitable tools are also required with which the fulfilment of functionalities and performance can be measured in terms of some appropriate



metrics. These test tools need to be developed, and criteria for their evaluation and approval are needed. In addition, a designation requirement for implemented methods and capabilities can be established for AI applications, for example by using the classification matrix for methods and capabilities in 4.1.2.

#### 4.3.2.3.4 Testability-by-design

Analogous to existing concepts such as “privacy-by-design” or “safety-by-design”, quality requirements for AI systems should also be taken into account in the design of the application.

The full life cycle of an AI system from the specification of the input data, the processing of raw data to training data and the representative modelling of a purpose-specific, domain-specific knowledge, right up to the application scenarios, must be considered. This also applies in particular to transparency requirements.

The concepts and standards of a testability-by-design for AI applications is a medium-term research topic. With regard to the quality characteristics mentioned at the beginning, at least the following fundamental research questions arise when using AI systems:

- How can an AI-specific FMECA (Failure Mode and Effects and Criticality Analysis) be performed?
- How will this have to be updated beyond the period of its development?
- For which purposes are an AI system and its AI components used? What are the resulting requirements for the testable design?
- Which AI models are used for the AI components employed? Are there standardized designs that are testable?
- Are people involved in the decision-making or prognosis by an AI component and if so, in what form? What are the responsibilities with regard to the input and output variables of the AI application?
- How are the AI models, implementation and training methods selected? What are the requirements for a testable design of the application?
- Which test methods are relevant, how can the AI component be tested if it has the appropriate properties?
- And how is the ongoing operation of this component monitored with regard to compliance with the purpose? What conclusions should be drawn for a testable design to simplify testing?

It is also to be expected that certain quality characteristics of AI systems will be easier to verify or their verification will only be possible if the corresponding requirements are already considered during the design and further development of the AI systems. Possible starting points are, for example, documentation of the development process, logging of (intermediate) results or interfaces for corresponding test tools (see 4.3.2.3.6).

#### 4.3.2.3.5 Test infrastructure for conformity assessment and certification

In order to be able to test the quality requirements formulated here, a testing infrastructure consisting of testing laboratories, technical inspectors and the necessary accreditation mechanisms and bodies is required. In particular, accreditation mechanisms should ensure that test bodies and testers have a sound technological understanding to perform these tests. When setting up the test infrastructure, the existing technical IT test infrastructure should be used as far as possible in order to develop marketable tests and establish connectivity to existing test methods. For the certification of persons, the competence of already established, accredited certification bodies can be extended with regard to methods and capabilities of artificial intelligence.

Certification may be based on a potentially updated variant of ISO/IEC 17024 [41]. In order to prove specific competences, further documents such as recommendations, regulations and further standards with regard to AI should be drawn up, extended and consulted.

The use of innovative, AI-supported testing services requires proof of existing professional competence of testers, technical experts, assessors and auditors in order to guarantee quality assurance. Apart from the validation of technical aspects, the potential for harm of an AI application should be assessable by qualified persons on the basis of ethical and legal principles.

#### 4.3.2.3.6 New test methods and new testing tools

##### Methodological approaches

According to a specification of the system to be tested, the test can be used, for example, with regard to machine learning procedures during training or on a fully trained system. This can be done by analyzing the input and output behav-

our of models to evaluate invariance, regularity and equivalence. Sensitivity analyses, for example, are suitable for this purpose. In the case of training-accompanied learning, a learning curve can also be tracked and intentionally evaluated in terms of declaration, error probability and adaptivity. For already trained systems, key performance indices can be included which evaluate criteria for suitability and exclusion of the AI for research purposes or a market. This should make it possible to determine, via interpretable quality characteristics, in which environment individual methods and capabilities of the AI can be used.

Using the LIME approach as an example, the goal is to explain systems based on machine learning [44]. Furthermore, models for the interpretation of learning mechanisms can be included for individual AI methods. For multi-layer neural networks, the methods “activation maximization” and “deep Taylor decomposition” are suitable [169].

Methods like LIME, Shapley [170], DeepLIFT [171] and QII [172] can often only be applied to structured data. Methods for unstructured data sets of data types such as text, images and audio are currently in an early stage of development.

A verification of the source code of AI-based systems is only possible to a limited extent using conventional software test methods. This includes statistical code analysis (GrammarTech’s CODESURFER), runtime verification (Java Pathfinder) or model checking (SPIN model checker) [173].

For different classes of neural networks different verifications can be specified which can be derived from different theories of logic and mathematics. These include verification procedures based on the satisfiability of formulas of Boolean propositional logic (satisfiability theories, SAT), the satisfiability of formulas of first-order predicate logic (satisfiability modulo theories, SMT), reduction to linear problems (mixed integer linear programming, MIP) and robustness of multi-layer perceptron networks (multi-layer perceptron, MLP). For SAT and SMT verifications the classical AI (symbolic AI) of automated reasoning is combined with ML. MIP is based on the logic and algebra of linear programming. Robustness studies of MLP apply findings from the theory of complex dynamic systems in ML [174]. Verification procedures for sub-symbolic AI systems and ML require new techniques which are extremely computationally intensive due to their parameter explosion (e.g. neural networks during autonomous driving).

### IT security tests for AI-based systems

An essential aspect of the tests for conformity assessment and certification are security tests, which are divided into static and dynamic tests. Dynamic security tests play a central role here, offering a wide range of methods and techniques. A brief overview is provided by overviews such as document ETSI TR 101 583 [175]. In this document is an enumeration and explanation of relevant methods and approaches for security testing, such as risk analysis and risk-based<sup>24</sup> security testing, functional testing of security functions, performance testing, robustness testing and penetration testing.

A number of techniques have been developed for the security testing of traditional software systems. These can only be applied to AI-based systems to a limited extent, if at all. Security tests for classical systems can partly be adapted for AI-based systems, e.g. the widely used fuzzing can also be used for AI-based systems in a modified form (see for example [176], [177]). In order to cover the security risks and attacks specific to AI-based systems, new techniques and approaches are needed that take AI-specific aspects into account, such as the relevance of training data. The techniques that address AI-specific security aspects are referred to as adversarial ML (AML) [178]. Coverage criteria are still a particular hurdle in security tests. There are a number of published metrics for this. However, a meta-study found only a low correlation between the existing metrics developed for AI systems and the robustness against attacks when these metrics were considered in tests [179]. An overview of existing techniques and metrics, including information on application, is currently the subject of the ongoing project ETSI DGS SAI-003 „Security Testing of AI“.

#### 4.3.2.4 National implementation programme for the Standardization Roadmap AI

The rapid spread and high complexity of AI systems is creating new technological challenges across industries. The dynamics prevailing in AI developments require a stable framework for action for all actors in research, industry and society in order to jointly use the available innovative power to shape the future and to promote the economic and social benefits of the use of AI systems in a converging manner. To operationalize those recommendations for action of the Standardization Roadmap AI that concern the technical

24 “Risk-based” in English.

requirements for AI systems, it was decided to propose a national implementation programme. The mission of this implementation programme is to develop such testing and quality assurance standards as central technical components of the action framework in a timely and needs-based manner, and to enable them to be updated in the future on the basis of economic and technical progress.

To encourage the success to the success of the mission, the programme will balance the short and medium-term demand from industry for operationalization of the technical aspects of the Standardization Roadmap AI and the long-term research needs on AI assurance issues.

Based on the results of the Standardization Roadmap AI summarized above, the programme pursues the following objectives:

- Development of the technological principles for a new generation of AI systems that are resilient and trustworthy “by design”.
- Development of expandable test criteria on the basis of established test technologies and those to be developed, using a uniform terminology.
- **Evaluation of these test principles in pilot projects with industrially mature, hybrid AI solutions in the course of a continuous improvement process with broad participation.**
- Derivation and development of reference architectures and test profiles for use cases and AI technologies with the aim of reducing testing efforts.
- Design and establishment of a test infrastructure for conformity testing and for certification at Federal level on the basis of existing test infrastructures.
- **Standardization of the test principles and their classification on the basis of existing standards and criteria. Establishment of the test standard at European level.**

The basis for the implementation programme is the joint programme CERTIFIED AI of the Fraunhofer Institute for Intelligent Analysis and Information Systems IAIS and the Federal Office for Information Security (BSI), existing international standards and specifications and their implementation in the field of IT security by the BSI, and the research activities of the German Research Centre for Artificial Intelligence (DFKI).

### 4.3.3 Standardization needs

The following information and needs concern the implementation of the Standardization Roadmap AI on the basis of the concept mentioned in 4.3.2.4.

#### NEED 1:

##### Implementation programme

At the centre of the entire testing system is the technical testing of requirements for AI algorithms, models, methods and data according to internationally valid standards. The development of such standards and specifications must be the focus of the implementation of this Standardization Roadmap AI. To this end, a national implementation programme should be established on the basis of the initiative mentioned above in 4.3.2.4.

#### NEED 2:

##### Relationship between technical requirements on the one hand and legal and ethical requirements on the other

Technical tests are crucial for the confidence and acceptance of the use of AI. Technical product properties must be recognized and sufficiently evaluated before their legal or normative permissibility for use can be determined. The implementation programme should initially enable the certification of technical requirements of industrially mature AI applications with justifiable testing effort. The conformity of non-technical requirements for AI systems can then be checked methodically separately using separate criteria catalogues.

#### NEED 3:

##### Embedding in existing test schemes and test infrastructures

In order to realize a timely implementation of the Standardization Roadmap AI in cooperation with research, industry, social bodies and governmental agencies, existing test methods and test infrastructures should be used. The development of criteria and methods must be based on existing, internationally valid test standards and test methods.

#### NEED 4:

##### Development of AI standards as a participation process

Within the implementation programme, the necessary test criteria and test methods for technical tests of AI solutions should be developed and tested in a broad participation process to gain trust and acceptance in industry and society. The good practice of the Standardization Roadmap AI should be reflected by conducting pilot projects for use cases with

the participation of industry, improving and adapting test methods, and gradually achieving standardization maturity.

**NEED 5:**

**Necessity of a management system standard**

The holistic implementation and interoperability at the level of the test concepts should be ensured in terms of standardization by embedding them in a comprehensive management system that takes into account organizational, technical and process-related test methods, as well as test schemes for different legal and ethical issues over the entire life cycle of AI systems.

**NEED 6:**

**AI Standards: Smart assistants for authorities and public agencies**

The fundamental democratic order is another strength of our society. It is based, among other things, on a large number of historically developed authorities, agencies and administrative processes that translate our values and standards into lived practice. This should now be opened for the Internet. The Online Access Act obliges the public sector to digitize many of the processes for citizens. At the heart of this effort will be smart assistants that establish a new form of human-machine interface or citizen-agency interface on the basis of AI. Smart assistants can have considerable advantages over the classic visit to the agency: proximity to citizens, 24/7 availability, speed, uniform quality, direct processing of digital documents, automation, cost savings, accessibility, ease of use for older people and people with disabilities, and much more. These advantages must be developed, preferably on the basis of uniform AI standards. Many of these will also affect the German language, which is particularly important. Why, for example, should not all town halls of a federal state have the same AI standard in order to productively set up a chatbot that explains opening hours, responsibilities and deadlines? There is an immediate need for action in this field. Division of labour between the authorities, clear responsibilities and an AI roadmap of “small quick steps” are crucial for success. Public sector AI standards make our community fit for the future and help to expand our values and processes digitally.

**NEED 7:**

**Need for research**

The development of high-quality test standards and test methods requires massive support from the relevant AI research community in Germany. For a complete operationalization of the recommendations for action there is still a considerable need for research in some areas, e.g. in the verification of neural networks. The research priorities must be worked on in parallel with implementation, and mutual synergies from research, implementation and standardization must be used in the best possible way.



## 4.4

# IT Security (and safety) in AI systems

Security and safety, especially IT security, are often considered spoilsports or even hindrances to innovation; however, history shows that security and safety have always been companions and promoters of innovation. It is not without reason that there are so many different standards, specifications and laws for security/safety, which have made comprehensive and trustworthy economic use possible. Without comprehensive security/safety and risk minimization, no car drives, no plane flies, no operation takes place and no house, bridge or road is built today. Security/safety have also been comprehensively regulated in more virtual areas such as electricity. People trust in safety-tested coffee machines or safe components in nuclear power plants and trained personnel. Innovations are not prevented, but their economic use is made possible by ensuring safety and security in use. This also applies to the application of information technology and especially AI. Of course, the consideration of safety and security always means additional effort and additional costs. Without safety and security, however, the costs and damage are no doubt much higher, see the recent case of production shutdown at Honda [180]. In addition, there are potential political dimensions, such as the massive cyber attacks on government agencies and businesses in Australia that became known in June 2020 [181].

Determining how much safety and security are required is a weighing of the effort required to accept a possible damage and can only be decided after the weighing (risk impact assessment) has taken place. This also and especially applies to the use of AI, since additional unknowns due to stochastic results and dual use possibilities are added. In order to be able to implement innovations economically on a broad basis in the market, trust must be created, for example by proving IT security.

#### Basic principles of IT security

The German term “IT-Sicherheit” is used ambivalently. It is therefore important to first clarify the three most important partial aspects when considering the topic.

“Sicherheit” can mean **safety**, which refers to the expectation that under certain circumstances a system will not lead to a state in which human life, health, property or the environment are endangered.

An IT system is functionally secure (safe) if its actual functionality corresponds to the desired, specified target functionality and the system does not assume any unauthorized states.

A functionally safe system is information secure if it only assumes such states that do not lead to unauthorized information acquisition or information modification. A functionally safe system is data secure if it only assumes such states that do not allow unauthorized creation, deletion, reading or modification of data objects. IT systems that are information- and data-secure are deemed to be reliable.

**Security** aims to prevent negative effects that a human or another machine can have on the AI module. Confidentiality, integrity and availability are the most important security objectives.

In addition to requirements for functional safety, such systems generally have special requirements for the confidentiality of information. The integrity of data (data protection) in connection with IT systems describes the control of a natural person – as a socio-technical system subject – over the disclosure of personal information and the availability of objects and subjects.

**Privacy/data protection (data security)** refers to the collection and processing of personal data according to the relevant regulations such as the EU General Data Protection Regulation. For example, persons in Europe have the right to have their private data adequately protected against IT attacks.

These three aspects – safety, security, privacy – are interrelated and can support each other, but also carry the potential for conflicting goals. For example, a high level of data protection can have a negative impact on the security objective of availability. In the case of the Germanwings crash in 2015, a “security feature” that protects the cockpit from terrorists (attacks) became a “safety problem” that threatens the lives of passengers. These relationships must be analyzed and taken into account during design and operation. Traditionally, this is done within the framework of a risk analysis.

Every AI system requires an individual analysis of its security and safety. Further research seems necessary in this environment, and the industry is recommended to develop the corresponding security levels. The applicable regulations, standards and specifications for ICT systems must be taken into account.

From an IT security perspective, AI systems are special IT systems to which the basic principles of information security apply without restriction.

In IT security research, an IT system is understood to be a technical system for storing and processing information. An IT system is closed if its technology comes from one source, if it is not compatible with other IT products, and if it is spatially limited. An IT system is open when it is networked, physically distributed and ready to exchange information based on standards. Open IT systems are usually not centrally administered, and their subsystems are heterogeneous. IT systems are components of socio-technical systems. They are embedded in social, economic and political structures and are used for a wide variety of purposes with overriding intentions. When considering IT systems, normative, legal and organizational rules and regulations, and questions of individual user acceptance and overall social acceptance can play a role.

IT systems process and store information that is presented as data. Data objects have the ability to store information and are created, deleted, read and changed by technical processes, i.e. by active subjects. Data objects, the information contained in them and the subjects for their processing are the assets worthy of protection within an IT system.

The discipline of IT security comprises all goals, procedures and measures to design, produce, operate and maintain information technology systems in such a way that a maximum of protection against operating errors, technical failure, catastrophic failures and intentional manipulation attempts is given.

#### 4.4.1 Status quo

In order to make proper use of the opportunities offered by artificial intelligence for the benefit of all those involved, one should know the risks and counter these with appropriate measures – a task of IT security.

AI solutions or systems are essentially complex systems of information and communication technology (ICT or more broadly known as IT systems). It is to be expected that AI systems will become or already are the target of cyberattacks. In its last survey in 2019 [182] the digital association Bitkom came to the conclusion that three out of four companies were victims of cyberattacks with damages amounting to more than 100 billion euros per year.

The existing IT security requirements for an IT system must be considered as the status quo when using AI.

#### 4.4.1.1 IT security standards and specifications

IT systems are already the subject of a wide range of standards, specifications, laws and regulations relating to security, IT security, safety and privacy with different histories. In addition, there are the standards and specifications for risk identification and treatment, some of which are independent and some of which are included. Their scope and diversity pose a challenge for companies and public authorities when using AI and its IT security. In addition, subject areas and industries which have so far used their own standards on IT security are converging as a result of increasing digitization.

The naming of examples of security-relevant standards and specifications in [Chapter 6](#) makes no claim to completeness, especially since AI systems are also used in industrial production environments (operational IT = OT) and for tasks with safety requirements.

Bitkom and DIN have also jointly developed a compass for a first insight into the subject matter [183]. There is also the DIN/DKE Standardization Roadmap for IT security [184], which could be extended to include AI topics. Standards that include the specifics of AI systems in terms of IT security are not yet available, but are partly under discussion.

Further research and harmonization for AI is part of the need for standardization.

#### 4.4.1.2 Laws and regulations

As IT security/cybersecurity has become elementary for critical infrastructures and companies, and also for consumers, due to increasing networking and digitalization, various regulations and laws have been issued:

→ At European level

- NIS Directive (Network and Information Security Directive 2016/1148) [185];
- GDPR, General Data Protection Regulation [95];
- LED Directive 2016/680 especially with regard to the processing of personal data by the police and judicial authorities [186];
- Directive on privacy and electronic communications [187];
- Cybersecurity Act [188];
- Machinery Directive [94];
- Product Safety Directive [189];

- and conceptual approaches (also include IT security/cybersecurity topics):
  - ethics guidelines of the AI HLEG [5]
  - White Paper AI of the EU Commission as a template for regulation [15]

→ In Germany

- IT Security Act [190] currently being revised as Version 2.0;
- Second Data Protection Adaptation and Implementation Act EU – 2. DSAnpUG-EU [191] (newly adapted federal data protection law), regulations at German Länder level;
- Telemedia Act (TMG) [192] for internet services;
- Telecommunications Act (TKG) [193];
- Law on the Federal Office for Information Security (BSI Law, BSIG) [194];
- Product Safety Act (ProdSG) [195];
- German Energy Law (EnWG) [196];
- various sector-specific regulations;
- the German Data Ethics Commission deals in its report [10] among other things with IT security with regard to AI systems.

Due to their importance for IT security of AI systems, the EU Cybersecurity Act, the EU White Paper on AI, the EU Machinery Directive and the EU General Data Protection Regulation are briefly presented below.

#### EU Cybersecurity Act

The EU Cybersecurity Act was adopted by the EU on April 17, 2019 [188].

The aim of the Cybersecurity Act is to establish IT security throughout the EU with uniform regulations and to strengthen systems, services and processes for information and communication technology (ICT).

In the future, ICT systems will be classified and certified according to defined “assurance levels” and their trustworthiness in 3 levels. The classification is based on a risk assessment with regard to the probability of a security incident occurring and its impact.

#### The assurance levels:

Level “Basic”: The basic risks for security incidents and cyber attacks are assumed to be low. At this level it is also possible for a manufacturer to assess conformity themselves and under their sole responsibility. Level “substantial”: Certified

products, services and processes should be able to withstand known cybersecurity risks. Level “high”: State-of-the-art cyber attacks can be fended off against attackers with extensive skills and resources. At this level, certification is only permitted by official bodies.

The 110 recitals of the Cybersecurity Act contain even more far-reaching criteria for risk assessment, objectives and basic requirements for cybersecurity and minimum components for the assurance levels. Among others, Recital 12 calls for “security by design”, Recital 13 “security by default”, Recital 41 “privacy by design” and Recital 49 “well-developed risk assessment methods and measurable security”. Certification is voluntary. The EU Commission will regularly examine whether cybersecurity certifications should be made mandatory, for example for companies in the energy, banking or healthcare sectors.

The Cybersecurity Act also provides for the implementation of ENISA and regulates its business activities. ENISA has been given expanded responsibilities and competencies, including the development of certification schemes for information security and the consideration of existing standards and specifications. In cooperation with national legislators and organizations, security experts, manufacturers of ICT products and users, the security criteria will be developed over the coming years.

Possible IT security risks from AI are not specifically described or considered. It is recommended to check to what extent additions might be necessary.

#### EU General Data Protection Regulation (GDPR)

The GDPR [95] strengthens and standardizes data protection in ICT systems for persons. Articles 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, for example from Machine Learning (ML) models that affect people, the following passage is crucial: “Where personal data [...] are collected, 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 throughout Europe have agreed on nine criteria:

1. assess or classify,
2. automatic decision-making,
3. systematic supervision,
4. confidential or highly personal data,
5. large scale data processing,
6. synchronize or merge records,
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 evaluation are relevant when using an AI where personal data are 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.

#### EU Machinery Directive

The EU Machinery Directive [94] regulates uniform requirements for machines and partly completed machines for a uniform level of protection to prevent accidents when they are placed on the market. In Germany, this has been transposed at national level as the Produktsicherheitsgesetz (Product Safety Act – ProdSG) and the Maschinenverordnung (Machinery Ordinance – 9. ProdSV) based upon it. The following requirements must be implemented (excerpt):

- The machine must be designed mechanically and electrically safe, and functional safety (e.g. safe control circuits) must be implemented,
- At the time the machine is placed on the market, it is safe and safe operation is guaranteed,
- Safeguards and protective devices of the machine cannot be bypassed easily,
- Conformity assessment procedures with risk assessment (§ 158 ff) are carried out,
- After successful assessment, the declaration of conformity is made and the **CE marking is applied**,
- Preparation of technical documentation and operating instructions that clearly draw the attention of the user and machine operator to the marked, existing residual risks.

Where AI components are installed in or for “machines”, the requirements of the Machinery Directive apply. There are no specific considerations regarding risks from AI and related measures. This circumstance could change in the coming years in such a way that besides safety, security aspects are also incorporated, which then also apply to AI.

#### EU White Paper AI

The EU White Paper AI [15] describes in 31 pages the basis of a possible general legal framework for the development and implementation of AI applications. To this end, the Commission takes up the recommendations and the seven key requirements of the “High Level Expert Group on Trustworthy AI”:

- human agency and oversight,
- technical robustness and safety,
- privacy and data governance,
- transparency,
- diversity, non-discrimination and fairness,
- societal and environmental well-being,
- accountability

One of the aims of regulation is to create an “ecosystem of trust”. IT security in terms of safety, security and privacy is reflected in key requirements 1 to 4. The White Paper notes that certain characteristics of AI (e.g. opacity, complexity, unpredictability and autonomous/semi-autonomous behaviour) make effective implementation of legislation difficult. An improved, risk-based<sup>25</sup> legal framework and its enforcement appears desirable to the Commission and should increase confidence in the security of an AI and thus its marketing opportunities. The application of the legal framework is basically „only“ planned for AI systems „with high risk“. Criteria are proposed for clarifying when a high risk exists. Also mentioned are possible measures and a possible obligatory determination of conformity, e.g. according to the „Cybersecurity Act“. For the other AI applications, the general rules should apply and these may be subject to „voluntary marking“ in the form of a trustworthiness seal of approval. Existing structures should be taken into account, both for governance and conformity assessment. The White Paper is currently under consultation.

#### Germany: German IT Security Act

The Act to Increase the Security of Information Technology Systems (IT Security Act, IT-SiG, 2015, Omnibus Law pertain-

<sup>25</sup> “Risk-based” in English.

ing to the BSI Law, EnWG, TMG, TKG) [190] has as its core objective the improvement of the availability and security of IT systems, digital infrastructures and services, as well as a better protection of citizens on the Internet. For critical infrastructures in Germany whose failure or impairment has a significant impact on the economy, state and society, such as energy and water supply as well as health care, it contains regulations on minimum requirements for IT security, duties of proof and reporting obligations. Currently, the IT Security Act Version 2.0 is being prepared, with which, among other things, the operational powers of the BSI are to be expanded and further parts of the economy are to be obliged to comply with the minimum requirements for IT security. Furthermore, it is planned to impose requirements on the trustworthiness of core components of the IT infrastructure used by operators of critical infrastructures.

The IT Security Act must also be considered for AI applications, but does not contain any special requirements for the use of AI.

The industry association Bitkom also offers an overview in its study from 2019 [197].

#### Report of the German Data Ethics Commission

Excerpt from the report of the German Data Ethics Commission [10]:

“Robust and secure system design includes both the security of the system against external influences (e.g. through encryption, anonymization, etc.) and the protection of humans and the environment against negative influences by the system (especially through a systematic risk management approach, e.g. based on a risk impact assessment). It must also include all phases of data processing and all technical and organizational components. Risks can arise not only from the technical design, but also from errors that human decisions in dealing with algorithmic systems bring. Since algorithmic systems and their embedding in an organization’s other information technology are not static, a management system is also required that checks and ensures the effectiveness of measures in the face of changing conditions, such as newly identified risks.”

The Data Ethics Commission for Artificial Intelligence considers, among other things, the security and robustness of an AI as an essential requirement, and presents 5 levels of criticality as a pyramid (see Chapter 4.1 to 4.3). In standards and specifications, however, a matrix is usually the common

and evaluated basis for a risk assessment. For this reason, the following representation by Krafft and Zweig [136], which was the basis for the criticality pyramid, was preferred.

With the help of a risk matrix (see Figure 21), based on the two characteristics (potential for harm through misjudgment and re-evaluation possibility), the application scenarios of ADM systems (algorithm decision-making) can be easily located, so that one can quickly get an initial overview of possible risks of the system. The risk is composed of the two components: the total harm to all individuals plus a possible superlinear total social harm. The individual regulation classes result in different transparency and traceability requirements, which are classified in Figure 21.

#### 4.4.1.3 Conformity assessment and certification for IT security

IT security can certainly be understood as part of quality, but it has its own extensive history with diverse roots in IT/OT security, safety and privacy, which must be considered in the context of AI.

IT security is connected with a proof (required by law or otherwise). Extensive recognized national and international auditing and certification procedures are already available for this purpose, in particular ISO/IEC 27001 [122], ISO/IEC 18045 [51], ISO/IEC 62443 [198]–[209] and BSI Grundschutz [184]–[186]. According to the Machinery Directive [94], conformity assessment procedures (CE marking) are available for products.

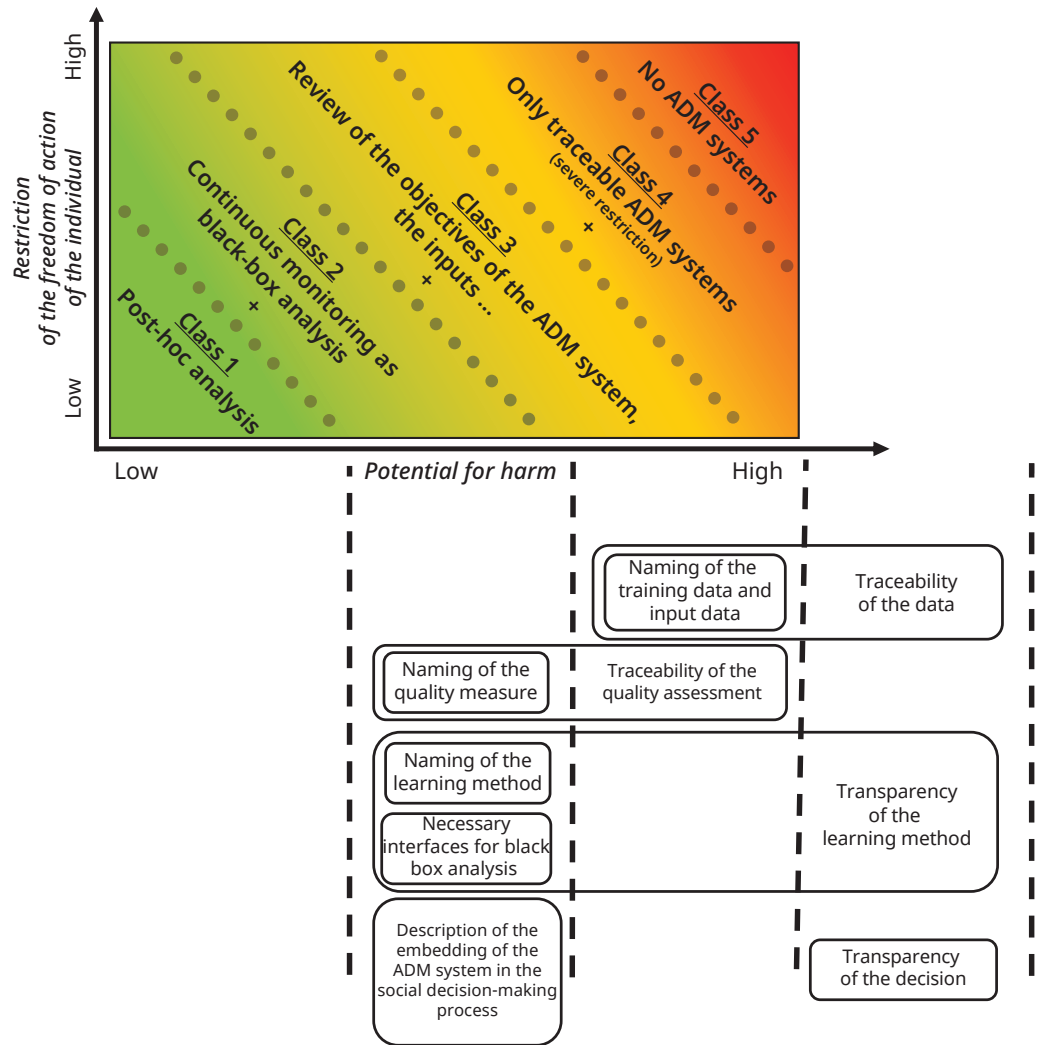
Within the framework of the EU Cybersecurity Act [188], further certification procedures are to be developed, taking into account existing procedures, which may also become mandatory.

There are also other conformity assessments and certifications for various industries, such as in the areas of mobility, health, supply or occupational safety. Further considerations on quality, conformity and certification can be found in Chapter 4.3.

#### 4.4.2 Requirements, challenges

Probably the greatest challenge for the use of AI systems by industry is the difficulty in establishing trust in (IT) security

**Figure 21:** Risk matrix with presentation of the transparency and traceability requirements according to the regulation classes [136]



and in the AI system. Among other things, trust can be created by inspecting and certifying the IT security of an AI system, which can be based on standardization.

IT security of AI systems is relevant in the office environment as well as in the industrial, operational (e.g. Industrie 4.0; IoT) and safety environment (e.g. automated driving or flying), although with different challenges and requirements.

For IT security, the well-known IT protection goals (Confidentiality, Integrity, Availability; CIA) form a basis, especially in the office environment.

In industrial applications, for example in production, the availability of the machine, the production, the device or system is in the foreground. A security update or a failure can have far-reaching consequences, as e.g. misconduct can lead to endangerment for humans or to considerable economic losses.

From a safety point of view, the protection of life and limb is the most important protection goal in every case, and extensive regulations must be observed when using an AI.

Just to consider these different perspectives for the use of an AI system, possibly simultaneously, is a challenge. Added to this are the supplementary requirements from existing regulations and standards.

Additional IT security risks and associated possible additional protective measures arise from the technological properties of an AI as a learning and changing IT system. This is due on the one hand to the great influence of the data on the training, testing and operation of an AI, and on the other hand to the type of data processing in the AI (e.g. with ML methods using neural networks). From an IT security perspective, there is a need for action to develop suitable protective measures for these additional risks.

The appropriate and required degree of IT security of the overall system depends on the purpose of the AI and its criticality and risk potential with regard to the probability of occurrence and impact of damage. Here there is a very wide range of possibly less problematic application scenarios (e.g. as a marketing tool up to AI controlled systems in mobility), which could mean a higher risk potential for human life due to your decisions. Therefore, the question must always be asked what expectations and requirements are placed on the AI component and on the overall system in which the AI component is integrated (e.g. in semi-autonomous vehicles, in video systems with pattern recognition, in IT systems for automated insurance policies) with regard to the identified hazardous situations.

In April 2019 the AI HLEG published ethical, legal and technical key requirements for trustworthy AI-based systems [22], including

- for technical testing of robustness against attacks, and of security, reliability and reproducibility,
- for the protection of data and their integrity, and
- for transparency and explainability of the algorithms ranging up to considerations of fairness.

There is a need for investigation and research to find out which additional risks and new requirements from an IT security point of view are to be expected for the creation and operation of an AI system, as well as its verification and certification.

Manipulated or deficient data or manipulated training can influence IT security. Sometimes even a certain fuzziness of the data used (e.g. minimal pixel change in video data) is sufficient to lead to misinterpretations. The handling of such a fuzziness has not yet been researched. However, this may influence the recognizability of whether a datum is (in the sense of faulty) true or false.

The environment, for example the availability of technical resources, also has a great influence on the security of an AI system. An attack scenario could be the manipulation of the resources to influence the latency of real-time AI and thus the result. There may be a dependency of the result quality of the used model on the available resources (e.g. computing power), hardware, infrastructure, interfaces and environment.

The IT security requirements for AI systems also depend on which actor/market participant is currently working with the AI. This refers to the possibly different requirements for man-

ufacturers, suppliers, integrators, operators and end users, from those who “only” have to deal with an AI software, an embedded system, an AI controlled plant/robot, an AI system consisting of various subproducts, or an AI cloud solution that is possibly networked with an AI IoT device, to the end user of an AI system. The application environment can range from uncritical to high or too high risk, whereby, for example, the AI software has not yet differentiated these fields of application in its origin at the manufacturer. An example would be AI-supported image recognition as a basis for diverse fields of application.

#### 4.4.2.1 Secure data

Currently, it can be assumed that in the near future, attackers will introduce falsified data into AI systems, whether to manipulate the results, divert resources from legitimate data sources or commit industrial sabotage. This could be countered with mechanisms that ensure that the algorithms operating on these data can identify and reject them.

The problem of detecting falsified data is not new, and machine learning essentially inherits this problem from the data on which it operates. Moreover, the process of adequately verifying the origin of data sets for training and feeding into algorithms and the associated issue of realistically assessing risk and liability may well delay the development and application of AI technologies.

A classical application of ML is the creation of predictions based on input data. However, the prediction of an AI system is based on the quality of the input data. If these are distorted or incomplete, errors can get into the AI, so the prediction is not reliable.

With machine learning, bad data is difficult to detect or remove. From a certain level of learning progress it is almost impossible to find out on which data elements which system internal decisions are based. “Forgetting” or “unlearning” is currently almost impossible. Data integrity and data quality are decisive quality characteristics for the success of ML systems.

If one considers the quality characteristics of data and metadata according to the definition of the Fraunhofer Guideline NQDM 2019 [90] (see also 4.1.2 and 4.3.2), all of them can have an impact on the security of AI systems in case of intentional or unintentional influence or insufficient quality. However, the quality characteristics of the data must

be considered differently depending on how it is used in training, testing, design or operation. Distorted training and test data can seriously affect the reliability of an AI system. An “unlearning” of distorted data is currently nearly impossible. Data in operation may have to be protected against manipulation at the physical level.

Especially under the aspects of information security (security) and confidentiality (privacy), the quality characteristics transparency and trustworthiness or accessibility and availability are to be highlighted. The latter are discussed in greater detail in the following (see 4.4.2.1.1).

Transparency and trustworthiness of data are important in any kind of data use and should be traceable, as they also increase confidence in data quality and data integrity. Data integrity and data quality are decisive for the success of ML systems.

Data integrity refers to the consistency, accuracy, trustworthiness and reconstructibility of data throughout its life cycle in IT systems. It includes measures to ensure that protected data cannot be removed or altered by unauthorized persons during processing or transmission. In IT security research, data integrity, data protection and data backup are essential requirements for reliable information systems.

Data quality is an essential component of data management because the quality of the data determines the credibility of the applications. This is, of course, especially true for data-driven technologies such as machine learning or big data analytics applications. Deterministic analysis and statistical data processing document and fix relationships between data elements. Expectations for the analysis of the data are hard coded.

Machine learning and especially deep learning independently generate and refine algorithms during the learning phase. To ensure the variation required for the exact development of the algorithms, deep learning requires sufficiently large and usually much larger data sets than conventional analytical applications.

Machine learning methods in general, and deep learning models in particular, require trustworthy training data sets. Sufficiently sound processes for cleansing the data are indispensable. The required data volume and the evolutionary methods in machine learning lead to fundamental questions such as:

- Where does the data come from? What system provides the data?
- How is the data accessed? Is the integrity of the data maintained?
- How are the data to be understood? Which relationships are there between the data? Are there dependencies and what type are they?
- How are and which data are used in analysis processes? How are data combined? How can the data be improved?

The quality of preparing data for an AI system is called curating. Developers should document the following properties to maintain transparency:

- the source of the data,
- the form of refinement (defining, collecting, selecting, converting, verifying) and enriching of the raw data into model or training data,
- the learning style (supervised learning, unsupervised learning, or other),
- the learning model used,
- the potential use of a special AI component,
- human participation (e.g. user feedback or labelling) in the decision-making processes within a processing operation,
- the institutions that produced the components of the AI system and that have decided on the selection, configuration, implementation and operation of the AI technology used,
- the curation of the data, the training and the selection of the models,
- the implementation of the AI algorithm, especially the rule-based instructions and decisions,
- the installation of test anchors and test agents.

#### 4.4.2.1.1 Protection of data, methods and measures

The processing and refinement of data in the development and operation of learning systems should take into account the nine criteria of technical understanding of the GDPR. There are basically three technical procedures to prevent data protection violations:

Data **encryption** is one way to protect sensitive information. As a rule, the processability of the data is then limited or even impossible. Homomorphic encryption methods can make operations on encrypted data executable, but this is often

too cost-intensive. With most other encryption methods, data must first be decrypted before operations can be performed.

But encryption also prevents publication of data even where it is necessary. One method of publishing without violating privacy is the **anonymization** of data. Anonymization distinguishes between three types of data: Identifiers, quasi-identifiers and sensitive values.

Identifiers are details with which a person can be directly identified, e.g. name details.

Quasi-identifiers are combinations of characteristics in order to make unique assignments. For the anonymization of data, certain rules and guidelines must be observed to ensure that no conclusions can be drawn about persons despite anonymization. However, attackers with sufficient background knowledge can deanonymize data. This risk can hardly be realistically assessed due to the range of possible relevant knowledge, with the consequence that anonymization is not necessarily sufficient to guarantee data protection.

Sensitive data (according to GDPR) means personal data containing, for example, racial or ethnic origin, political opinions, religious or ideological beliefs or trade union membership, as well as the processing of health data, genetic data, or biometric data for the clear identification of a person.

The concept of **differential privacy**, on the other hand, can guarantee data protection by providing, in principle, a statistical guarantee that the data of individual persons have no effect on the result of certain queries.

The basic principle is that the privacy of a person is guaranteed precisely when the result of a data query does not depend on the data of a single person. This requires functions that can answer database queries while ensuring that data protection is not violated. The original data are provided with “noise” or modified during a query. The modified data cannot be distinguished from the original data and statistical relationships are not distorted. However, the effects on learning processes need to be examined more closely.

#### 4.4.2.1.2 Security and trust in authenticity, integrity and quality of data, methods and measures and discussion of the approaches

AI systems increase the complexity of IT security as compared with “normal” IT systems. This is due both to the software tools used, e.g. machine learning and neural networks, their interaction with and dependence on the environment, and to the much more intensive importance of the data used for training and in productive use.

In view of expected future attack scenarios, resistance to attackers is a key requirement for cyber-physical value chains. The following four example (see Table 6) approaches represent possible solutions to establish trust in the authenticity, integrity and quality of a specific ML data label (machine learning) and should be considered when designing and developing future ML applications to increase the resilience level:

**Table 6:** Four example approaches to establish trust, integrity and quality of an ML data label

Method/ Approach	Description	Security against manipulation
1 Reputation systems	Correlation of events and feedback results about the identity subjects who created the data sets	low
2 Algorithmic analysis	Analysis of output data sets based on machine learning	low to medium
3 End-to-end data provenance	Authenticity analysis to verify the concrete origin of the data and the integrity of the data chain	medium
4 Identifiable data provenance with scoring mechanism	Analysis of the authenticity and integrity of the data’s origin and evaluation of the entities involved based on their life cycle certificates and historical events, where available	high

Other hybrid models combining these methods can also be developed.

In principle, however, all analyses are more robust when combined with a scoring mechanism based on reputation/ data origin. For this reason, a discussion on how to develop better global reputation systems proves to be inevitable (also in order to make reliable data securely available to more players in the market), since the use of MLs alone does not adequately verify that input data or labels based on them have not been falsified by similar MLs.

### APPROACH 1: REPUTATION SYSTEMS

**Centralized data reputation systems** can be mapped on a large scale on monolithic platforms. Typically, a marketplace has a native reputation system that works independently and is abstracted from unique personal identities. The lack of robust verification and assessment mechanisms allows participants to manipulate these evaluations.

The integrity and authenticity of data cannot be verified without access to the identity registers of a central platform, even assuming the integrity of the content of this ideally well-managed register.

**Decentralized data reputation systems** and pseudonymized, token-curated registers, mapped on a block chain infrastructure, would be able to verify unique digital identities for all participants in an open system and aggregate reputation data across all platforms where the data subject has agreed to be correlated for reputation purposes.

However, this “web of trust” approach to reputation disclosure is still at an early stage and its own unique attack vectors still need to be tested in practice. Until such systems are fully developed, such decentralized reputation scores can be used as one data source among many for a probabilistic, hybrid scoring model.

### APPROACH 2: ALGORITHMIC ANALYSIS OF THE OUTPUT DATA AND ITS LIMITATIONS

Another approach is the analysis of the output data of an IoT device or an algorithm with ML algorithms. The following techniques (see [Table 7](#)) are used to determine whether a given data set is falsified or genuine:

**Table 7:** Techniques for determining truthful data sets

Output vector	Description	Example Image processing
Object Features	Analysis of selected features or locations of an object where algorithms that generate the falsified object typically fail	Visible artefacts at the interface of hair and body in a human image
Format Features	Analysis of content related to specific format characteristics	Falsified images tend to have smoother textures
Neural Monitoring, a.k.a. Reflexive Monitoring	Analysis of neurons and layers of the network that are activated during identification/processing of real and falsified images	Test how other advanced algorithms react to previously sorted authentic and falsified images

Static criteria for all three of these analysis vectors can be provided manually, but because enemy networks are trained on historical data, they quickly overcome any analysis. This leads to all three methods becoming three separate fronts in an “arms race” between learning algorithms, where none of the above methods can provide a final, resilient solution. Rather, they are subject to a circularity that opens possible attack vectors since all three methods relate to the use of machine learning to identify by-products of simpler or older machine learning methods, in an ongoing process that is never complete.

### APPROACH 3: END-TO-END DATA PROVENANCE

Today, the traceability of data origins is only partially given. As long as the data does not originate from its own specially controlled data silo, it is not possible to certify that purchased data are completely secure against forgery, since origins and data traces that cannot be verified from the outside are even easier to forge than the data itself.

If the data comes from your own trustworthy sources, there is no question that an efficient AI system can be set up with it. However, only a few players on the market have their own corresponding data volumes or have (secure) access to them.

End-to-end data provenance is basically the cryptographic signing of all data from a data source. This creates the ability to trace a data packet back to its origin, to the very device that first made a measurement or registered an event. Only through the identifiability of the data source can the prove-

nance of the data flow form the basis for verifiable assertions. By signing the data packets, a data chain<sup>26</sup> is created which makes it possible to assess the trustworthiness, reliability or risk metrics of the data source.

#### APPROACH 4: IDENTIFIABLE DATA PROVENANCE AND GLOBAL ASSESSMENT

Only a reputation system that is based on a neutral and complete audit trail can adequately assess the trustworthiness of learning systems.

Here, a scoring model “assesses” or estimates the risk of using data chains by assigning relative values of trustworthiness or validation to all unknown actors. Here, actors and agents are scored in a system based on their historical familiarity or accuracy.

In order to create a trustworthy environment for ML data, the first step is to cryptographically sign all output data from a data source (Approach 3). Only through the identifiability of the data source can the provenance of the data flow form the basis for verifiable assertions and certificates about the data flow itself as well as for reputation mechanisms.

The second step involves anchoring the resulting data chains from verified sources, in a decentralized identity meta-platform. This platform provides a public key infrastructure with which publicly anchored data identities can be created. In this way, the data genesis and each transformation event could be electronically signed, allowing the subsequent verification of any data flow and an estimation of the trustworthiness of the output data of a machine learning algorithm. These assessments can be made directly from the data source and/or indirectly using public/open registries and reputation systems.

An evaluation algorithm can request a kind of “lifecycle credential” from each anchored data flow and thus reflect or assess the general, aggregated trustworthiness, data flow

provenance<sup>27</sup> and accuracy of a machine learning data label. The public anchoring of data reputations is a prerequisite for adequate objectivity in a mature reputation system. This function could be well represented by large institutions with public tasks.

#### 4.4.2.2 Robustness and security against attacks – attack vectors and defence mechanisms

Based on the holistic approach of IT systems with regard to their security, the entire model infrastructure should also be considered for AI systems in addition to security and data protection. These essentially include three protective measures:

- model authentication (e.g. through a role-based restriction of the use of the model on a specific terminal device and for a predefined duration; this is particularly important in the case of so-called distributed learning (federated learning)),
- secure model ownership and distribution (e.g. by encrypting the model and maintaining secure transmission paths),
- model verification (e.g. by means of a suitable comparison to reference models in order to detect anomalies with regard to the expected mode of operation).

The measures outlined here already cover a large part of the spectrum of attacks (including model theft and denial of service attacks) that can have serious economic consequences.

##### 4.4.2.2.1 Adversarial Machine Learning (AML)

When considering the robustness of AI systems, the term AML is of central importance. This describes a current field

26 A data “chain” is any cryptographic data structure that “chains” signed data objects together (with unidirectional or bidirectional “connections” between entities), thus creating a navigation method for comprehensive data flow provenance and verification. Data flow provenance enables the end-to-end integrity of each data flow object and its transformations to be verified.

27 By “data flow provenance” is meant a mechanism for tracking data points and the handling of this data by a processing system that registers every transformation to these data points. [This includes flows with multiple sources, collective sensor fusion and processing by machine learning algorithms. Comprehensive data flow provenance involves not only tracking the retention of data, but also checking the end-to-end integrity of each data flow, including all transformations (additions, deletions, modifications, combinations and ML processing)].



of research<sup>28</sup> dealing with the security aspects of ML. In a narrower sense, only possible attack scenarios are discussed, but in a more general sense it also covers the development of appropriately hardened (robust) models, detailed analyses of defence mechanisms and the evaluation of attack-specific consequences.

#### 4.4.2.2.2 Attack vectors and defence mechanisms

The following is an overview of both attack vectors (see [Table 8](#)) and defence mechanisms. It does not claim to be exhaustive, but offers an introduction and orientation for developing suitable practical solutions to meet security-related challenges in relation to machine learning processes by establishing uniform terms (terminology and taxonomy). The explanations are deliberately kept so general that they are independent of the specific learning paradigm. Only at individual points are concretisations made to this effect. Learning paradigms are, for example, supervised, unsupervised or reinforcing learning. Especially for reinforcing learning, whenever there is talk of influencing “data”, the environment of the system (in this case, in the common linguistic usage: an “agent”) should also be considered.

Basically, the components of ML systems can become the target of attacks using different techniques and the existence of certain knowledge levels. The goals result from the typical structure of an ML system, namely the physical domain and its digital representation

- of input sensors,
- of the ML model itself, and
- of the outputs or physical actions.

The state of knowledge can be described by progressively inclusive amounts of information available to the attacker.

In the simplest scenario, this person has no knowledge of the ML system, but can only make inputs or, possibly, also tap outputs. With regard to the strength of the attack it is decisive to what extent and in what form (direct output, probabilities, final classification) input-output pairs can be tapped.

In addition, the attacker may know information about the model family (e.g. a neural network, an ensemble of decision trees or a hybrid system) and its concrete architecture.

Further detailed knowledge, for example in the form of the parameters or even the learning algorithm (e.g. the optimization procedure and the loss function) including its hyperparameters finally form the highest level of knowledge. Usually in this context we talk about black-box, grey-box or white-box attacks – depending on how far the knowledge of the attacker reaches.

The techniques are distinguished according to their use during the training or operational phase of modelling or operation. The former aim to influence the data, the learning algorithm or the model, whereas the latter do not do so, but generate new inputs for the model that elude the intended functioning of the model (classification, regression, behaviour of an agent).

The first question to be answered when looking at **attacks during training** (poisoning attacks) is the extent to which the attacker can manipulate the data, i.e. add new data, delete data or modify existing data.

For the latter it is also decisive whether the attacker can only influence the inputs or also the corresponding outputs, i.e. complete input/output pairs. A data distribution learned in the sense of the attacker can, for example, lead to a reduction in the accuracy of the classification, the installation of a back door or the targeted guidance of an agent.

In the case of **attacks during the operating phase**, the training data and the model itself remain unaffected. Instead, these attacks design new test data that e.g. evade the intended classification (evasion attacks) or that can be used to collect information about the training data and/or the model (model extraction attacks).

<sup>28</sup> Terms that are already established in the research literature, which is mainly written in English, are adopted here without translation, since it is to be expected that they will also become established in the German-speaking world.

**Table 8:** Attack vectors

Attack vector	Characteristics
Evasion Attacks	Usually solve a limited optimization problem that looks for input examples that maximize the loss function of the model with the least possible deviation from regular training or test data. Mostly single or multi-step (iterative) gradient-based methods are used, more rarely also methods without using gradients. The latter usually require the corresponding output probabilities in the case of classification systems.
Model Extraction Attacks	Directed at the model, often intend to replicate it for selfish purposes or to generate evasion attacks using the substitute model. For the aggregation of input-output pairs a sufficiently unrestricted interface to the model is required. If information about the data is collected in the course of a model extraction attack, statistics about the distribution of the training data can be collected or the training data itself can be extracted (model inversion attacks). It is also possible to determine whether certain data points belong to the training data set by means of suitable queries (membership inference attack).
Influencing the model output	With complete or only partial knowledge of the respective decision paths. What is meant, for example, is a positive classification result in the sense of the attacker (e.g. in the assessment of creditworthiness) based on the knowledge of the branches in decision trees, the creation and evaluation of so-called saliency maps (e.g. for neural networks), the exploitation of the reward function in reinforcing learning (reward hacking), but also already the knowledge of a bias present in the training data and therefore most likely learned from the model. The partial knowledge about the coming about of the decision is useful for the attacker in the sense that e.g. the occurrence of certain categories can be excluded in the result of a classification.

According to the differentiation of the attack vectors, a differentiated consideration is useful for the **defence mechanisms**, depending on whether they are effective against attacks during the training or operational phase. It should be noted at this point that the safeguards often limit the performance, e.g. the accuracy or the inference time of the ML system and therefore a trade-off must be made between the two.

The containment of poisoning attacks can, in addition to general provisions regarding data storage and data provision,

be achieved by monitoring the inputs. Thus, manipulated data can be identified by means of suitable statistical methods under certain circumstances even before the start of the training and the training data set can be cleansed of them. In addition, there are approaches that can detect differences in the sequence of legitimate or manipulated data sets even under observation of the model itself, for example its learned parameters. A variety of potential procedures have been developed to mitigate attacks during the operation phase. The following list contains a compilation of examples with corresponding explanations:

- The training data set is supplemented by input/output pairs created by a specific attack (adversarial training)
- The attacker is denied access to usable information about the gradients of the model (gradient masking)
- Starting from a certain model, a new, less complex model is trained, whose decision boundaries are smoother (defensive distillation)
- The outputs of an ensemble of models are combined (ensemble methods)
- Randomness is introduced on the training data set or within the model (e.g. in the form of Gaussian noise) in order to reduce the information gain with a fixed number of tapped input-output pairs
- To make manipulations of the inputs ineffective, they are transformed in different ways, e.g. by compression or smoothing operations (feature squeezing)
- Inputs are moved near the nearest data points in the training data set using an upstream model (reformer)
- The ML system is only trained and operated with appropriately encrypted data sets (homomorphic encryption)

At this point, however, it should be emphasized that many of the mentioned approaches, e.g. gradient masking, can easily be circumvented by using a substitute model as a result of an extraction attack.

In conclusion, it can be said that attacker and defender are in a dynamic interplay, i.e. there is an interlocking of the considered attack vectors and defence mechanisms. In the case of ML systems, for example, this can be quantified using so-called security evaluation curves (accuracy of the model vs. strength of the attack). According to the current research consensus, a corresponding defence mechanism can be constructed for every conceivable attack vector, but an adaptive attack (in the simplest scenario already the same attack vector, but with changed parameters) can always be constructed for it, rendering the protection ineffective. For all further research efforts, this initially means that all new defence

measures must be evaluated comprehensively and carefully against adaptive attacks. Although some of these already exist, all guarantees made so far are based on insufficient metrics. Especially in the field of image processing there is no metric that can reliably distinguish legitimate from manipulated input in a manner congruent with the visual perception of a human being.

This approach can lead to short-term progress in understanding the mode of action of attack models and the vulnerabilities of ML systems. In the long term, however, this is not satisfactory from an IT security perspective, and the focus must inevitably be on developing more formal robustness guarantees. There is a need for research and standardization to be able to offer more reliable and more IT security in the future.

#### 4.4.2.3 IT security risk assessment

AI can be understood as a singular product (software for speech recognition) or as embedding in a system with different components (e.g. part of an online platform of an insurance company or part of the control of an industrial robot). The AI system in turn is part of an environment, has interfaces to it (e.g. knowledge database or video sensors as data source), which affect the AI. On the other hand, the AI has an impact (decision on insurance coverage or how the industrial robot works) on its environment. In addition to the many positive results, the effects can also involve risks of varying degrees of severity, including IT security.

The IT security risk assessment for AI concerns safety, security and privacy aspects (see [Chapter 4.4](#)) and there are several possible approaches.

For example the meta-model as in DIN SPEC 92001-1 [87] (see [Figure 14](#)) can be helpful. This recommends identifying security requirements (safety, security and privacy) for the three pillars “Functionality & Performance”, “Robustness” and “Comprehensibility” and classifying the respective risk.

For the sustainable IT security of an AI system, the entire “life cycle” must be considered. Each stage involves potential risks that should be analyzed and assessed. This begins with design and development and extends from testing, training, distribution and operation to decommissioning. The AI model, e.g. with machine learning, and the data for testing, training and operation, the product, the overall IT system and inter-

actions with the environment must be taken into account in the risk assessment. IT security management, or an AI system for safety, security and privacy such as in ISO/IEC 27001 [122] and ISO/IEC 27005 [210] can provide great support. To what extent the standard can or should be extended for AI systems would have to be examined.

EXAMPLE: An intelligent, digital video camera with an embedded trainable AI-based analysis module should already consider IT security for safety, security and privacy in the design and development stage if it is to be used as an intelligent end device (IoT) in different application scenarios, e.g. in self-driving cars or as a surveillance camera in sensitive access areas – areas with higher risk.

The protection of the data used, their quality and trustworthiness are of great importance, as they have an intensive influence on the function and results in testing, training and operation. The task of IT security is to ensure that neither the data nor the model, the system or the environment become the target of a successful attack during the test, training or operational phase. In order to be able to take preventive measures, a risk assessment is a mandatory requirement.

#### Variety of laws, standards and specifications for risk assessment for IT security

A wide variety of approaches, recommendations, models, procedures, standards and laws are currently available for risk assessment and classification, both in Germany and at EU or international level. Here, too, it is recommended to test and assess AI systems and to develop a practice-oriented common basis. The following are just a few examples, not all of which are specific to IT security and/or AI:

- The **German Data Ethics Commission** [10] recommends five levels of criticality of applications with no or low to unacceptable potential for harm and 7 evaluation criteria:
  - The dignity of humans
  - Self-determination
  - Privacy
  - Security
  - Democracy
  - Justice and solidarity
  - Sustainability
- The **EU White paper on AI** [15] describes two risk levels, i.e. AI with “high risk” and the other AI applications. There are references to high-risk AI applications, such as biometric remote indication in public places. The following criteria are suggested:
  - Human agency and oversight,

- Technical robustness and safety,
  - Privacy and data governance,
  - Transparency,
  - Diversity, non-discrimination and fairness,
  - Societal and environmental well-being, and
  - Accountability
- **The EU Cybersecurity Act [188]** (see 4.4.1) names three levels of assurance, Basic, Substantial, and High.
- **The Machinery Directive [94]** (see 4.4.1)
- **ISO/IEC 27001 [122] and ISO/IEC 27005 [210] Risk management IT Security** refer to ISO 31000 [93], but supplement it with more concrete points (excerpts):
- “the strategic value of the business information process;
  - the criticality of the information assets involved;
  - operational and business importance of availability, confidentiality and integrity;
  - stakeholders’ expectations and perceptions, and negative consequences for goodwill and reputation; Additionally, risk evaluation criteria can be used to specify priorities for risk treatment.”
- **IEC 61508 [79]–[86] and IEC 61511 [211]** designate 4 safety levels or Safety Integrity Levels (SIL) for functional safety, which are used to assess E/E/PE systems with regard to the reliability of safety functions. The evaluation criteria here are the danger to life and limb via the extent of damage and the probability of occurrence.
- **ISO/IEC 15408 [48]–[50]** lays down
- seven Evaluation Assurance Levels (EAL, see 4.1.2.2.2 “Common Criteria”) for the trustworthiness in the security performance of an IT system/product, via
  - eleven defined functional classes (e.g. security audits, communication, cryptographic support, identification and authentication, the protection of user data and security functions and Target of Evaluation (TOE) access), and
  - seven organizational classes for delivery and operation, development, quality of handbooks, functional tests and vulnerability analysis.
- **DIN SPEC 92001 [87]** classifies assessment criteria into three quality pillars: functionality & performance, robustness, comprehensibility, and two risk levels: low and high.
- **ISO/IEC 23894 AI Risk Management** (in preparation) lists possible assessment criteria (excerpt)
- Security
  - Privacy
  - Robustness
  - Availability
  - Integrity
  - Maintainability
  - Availability and quality of data
  - AI expertise
- **ISO 31000 [93]** Risk management guidelines
- The publication “From Principles to Practice; An interdisciplinary framework to operationalise AI ethics” [123] published by **VDE and the Bertelsmann Foundation** and other authors presents a risk matrix with 5 classes of AI application areas with risk potential. The classes range from “no ethics rating required” in class 0 to “prohibition of AI systems” in class 4.
- **Roadmap SafeTRANS “Safety, Security, and Certifiability of Future Man-Machine Systems”**, is an example of how security and safety are interlinked (see 11.3)
- **GDPR** (see 4.4.1) with the challenge of data protection impact assessment (risk) at “high” risk and identifies risk areas, e.g. health data.
- **ISO/IEC 29134 [212]** Data protection impact assessment
- **Various protection level concepts for personal data**, e.g.
- of the State Office for Data Protection of Lower Saxony, or
  - the Independent Data Protection Centre of Saarland, or
  - the Standard Data Protection Model (SDM) of the Conference of Independent Data Protection Authorities of the Federal Government and the States (DSK).

AI IT security requires security criteria and risk assessment for the secure product or system, as well as for a secure life cycle. The criteria and risks are also quite different depending on the application (use case) and actors, such as the manufacturer of the AI product, testing, training, configuration, installation, integration, operation and use of the AI system. The learning systems and data used require additional attention and possibly additional criteria.

The existing laws, standards and specifications offer different criteria and assessment benchmarks. This is a challenge for practical implementation in terms of the economy, and a basis with common criteria and assessment standards for IT security and risk assessment would be helpful. These could then be deepened and supplemented for each sector.

### 4.4.3 Standardization needs

The following topics have emerged as concrete recommendations for action for standardization, research and the public sector:

#### 4.4.3.1 Basis for standardization

##### NEED 1:

##### **Research/examination/evaluation of existing standards, conformity and certification procedures and existing laws**

For IT systems there are already various standards, specifications and regulations that can or must be considered for systems with AI. AI systems are increasingly used in industrial production environments (operational IT = OT) and for tasks with safety requirements. The first step should be an adequate research, examination and evaluation of existing standards, conformity and certification procedures and regulations for IT security and risk assessment in order to extend them with AI specifics and, if necessary, with documentation requirements. Proposals for harmonization and consolidation are also recommended.

##### NEED 2:

##### **Recommendations for actors/market participants**

Manufacturers, marketers or users of AI systems could be supported with standards and concrete recommendations for action in such a way that even those with less technical expertise can implement suitable IT security. With AI systems, the assignment of risk, criticality and trustworthiness may be difficult. For example

- a more universal AI is developed, pre-trained and brought to market by the manufacturer (pattern recognition by video).
- Further, this AI is trained/adapted/customized/installed by an integrator for more specific applications and marketed as a separate product (e.g. road sign recognition).
- This product is purchased by a company, possibly installed, further optimized/trained (e.g. in a vehicle) and made available to end users (drivers) as a complete system.
- This involves transferring data from the vehicle to a cloud service provider, which hosts the AI software in its data centre.

This results in complex questions regarding IT security, risks and related liability issues. Support could be provided here

within the framework of standardization, for example through practical instructions for action.

##### NEED 3:

##### **Development of supplements/adjustments in risk management**

For the risk assessment of IT security with regard to AI systems, it will probably be necessary to develop further standardization content. Within the IT security management standards family ISO/IEC 27000, a “separate” AI-specific standard ISO/IEC 2700x or a supplement to the existing ISO/IEC 27005 (risk assessment) [210] could be developed for AI systems. The appropriate procedure is to be considered.

#### 4.4.3.2 General framework for IT security

##### NEED 4:

##### **Combine criticality levels and IT security**

The levels of criticality proposed by the Data Ethics Commission should be examined with regard to their use for IT security, and as precise provisions as possible should be set out in standardization.

##### NEED 5:

##### **Define IT security criteria for training methods**

In terms of IT security, there are currently no clear criteria for the learning of AI systems. This gap in IT security criteria for training methods can lead to unintentional intrusion by external parties, use of malicious data or other manipulation.

##### NEED 6:

##### **Create explainable AI**

It is necessary to define the relevant aspects for transparency. This includes the two terms “traceability” and “verifiability”. Here the goal is clear: an explainable AI system.

##### NEED 7:

##### **Define controls for IT security**

For the standardization and testing of AI systems it is necessary to define appropriate measures (controls) for IT security (enterprise IT security, OT security, safety IT security; privacy). These are helpful for implementation and can be tested and certified.

##### NEED 8:

##### **AI security by design and AI security by default**

Effective IT security and data protection in an AI system must already be considered and designed holistically in the first

step of development (“design”) and form the basis of the functioning (“default”). Therefore, appropriate engineering requirements are absolutely necessary for standardization (“security by design and by default”) and could be included as a supplement in existing standards for secure software development.

#### 4.4.3.3 Data

##### NEED 9:

##### **Verification of the provenance and protection of data**

The protection of data against manipulation in an AI system is of high relevance due to the specific characteristics of AI systems. Mechanisms must be implemented to ensure that falsified data introduced (by attackers) into the operating algorithms can be identified and rejected by them – this is where research is needed. The clear traceability and/or verification of the data origin and use should be given and supported by methods. Technologies such as block chain or other cryptographic procedures could be considered here.

Based on this, it is necessary to clarify how (training) data should be handled if they are collected or used outside the later field of application. Risk assessment should consider unintentional and intentional influencing possibilities and weighting of (training) data (online data and offline data). In this context, the design and import of input data, the selection and origin of test and training data, secure and correct data in operation and output data should be a field of action.

##### NEED 10:

##### **IT security of training data**

Data and the training/operational models should be verifiable. For the training itself a specified semantics or semantic context is needed. For example, there are different types of training, such as “individual training” or several AI systems in a “group training”. The training concept should be integrated into IT security, examined for relevant risks and provided with possible protective measures.

#### 4.4.3.4 Learning systems

##### NEED 11:

##### **Define IT security criteria for learning systems**

Core elements of artificial intelligence are the learning systems (ML, Deep Learning) and their IT security. New IT security investigations and specifications are required here.

Any knowledge of influencing variables and risks must be taken into account in definitions, specifications and weightings that are generally related to IT security. Security in hybrid AI systems (knowledge-based and data-based approaches) is another field of action.

For the new requirements on criteria for learning systems and their components, it should be determined and, if need be, researched, which security controls, test procedures, auditing and certifications are necessary.

##### NEED 12:

##### **Verifiable identity of AI algorithms**

To strengthen trust in AI systems, AI algorithms must be provided with a verifiable identity and their function and mode of operation must be recorded in documentation. If possible, results should not only be shown as a probability value of a result class as the basis of a decision, but as a confidence interval.

#### 4.4.3.5 Research topics

##### NEED 13:

##### **IT security metrics for learning systems and adversarial machine learning (AML)**

In the field of AML, especially with regard to applications in practice, further research and development is necessary, since the analysis of the IT security of any AI system requires a complex, individual analysis of its security.

Among other things, the development of meaningful metrics for the evaluation of robustness should be aimed at, which can, in the long term, be used as a basis for standards and specifications to enable a comparability of robustness.

The research and development of a general taxonomy for AML is recommended which lists both attack and defence procedures that should be considered when developing models. In a further step a “tool box” could be made available which contains attack vectors and which can be applied to existing trained systems. In particular, the merging and standardization of already existing tools, defence mechanisms and robustness concepts should be considered. Standardized automated procedures would also be conceivable and helpful.

**NEED 14:****Impact of availability of resources**

The availability of resources, such as processor power, impacts on the IT security of an AI system and is relevant when considering different phases. First of all, this is the design time during development, then the testing and training time and then the runtime, i.e. during operation. In all phases a lack of resources can lead to faulty/incorrect results. This could also be an attack vector in terms of IT security. Here, research should be used to investigate a simulation of possible attack scenarios on the resources and working methods of the IT security of an AI system and to create a collection of attack vectors. These simulations in turn require a variety of (training) data (e.g. face recognition, rotation of faces, background of faces, skin colour, etc.).







**4.5**

## Industrial automation

Artificial intelligence (AI) is an important and essential key technology for maintaining Germany’s economic performance. In particular, AI has a particularly high potential to sustainably design workflows and processes in the manufacturing industry – i.e. Industrie 4.0 [213] – 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.

In English and increasingly also in German-speaking countries the term “industrial artificial intelligence” or “industrial AI” is also being used and covers all fields of application of artificial intelligence in industrial use [214]. This topic was structured within the context of the work of the project Industrie 4.0, which was initiated by the German Federal Government in 2015. Based on existing value-added processes of the manufacturing industry [215], corresponding future application scenarios defined [216]. The application scenarios cover a wide range of applications, such as order-driven production based on dynamic value-added and supply networks, versatile factories that enable the flexible adaptation of a factory’s production resources, smart product development, and much more. These application scenarios provide the basis for further refinements and analyses to derive possible research and standardization needs.

An important role in the digital transformation is ascribed to the digital representation of physical reality, the “digital twin”. In order to ensure interoperability within a digital ecosystem, the Plattform Industrie 4.0 is working with all participating institutions to develop the specification of the administration shell as a digital image of every relevant object (asset) in networked production. An administration shell stores all essential properties of an asset such as physical properties (weight, size), process values, configuration parameters, states and capabilities. The administration shell is not only an information store, but also a communication interface through 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 represents an important basis for the application of artificial intelligence for Industrie 4.0, as it allows a uniform access to data and metadata of relevant assets and makes them available in a structured data format.

Further applications of artificial intelligence are considered in the context of Industrie 4.0. In addition to autonomous intralogistics (see also the Chapter 4.6, Mobility and Logistics), industrial image processing and image recognition, as well as the improvement of the interaction and integration of humans and machine are taken into account. On the one hand, through the use of new interaction mechanisms, such as speech and gesture, new display options, such as augmented reality, and the strengthening of collaboration, such as through collaborative robotics. In these cases, AI technologies are used intensively throughout. Standardization aspects specific to industrial automation are currently being investigated. As part of the work of Working Group 2 of the Plattform Industrie 4.0, the influence of AI in selected application scenarios has already been examined in detail [23], [24]. Known application examples for AI in Industrie 4.0 include predictive maintenance, whereby the lifetime and necessary maintenance time of components are predicted based on symbolic models and collected operational data. In addition, the continuing monitoring of the production processes, the prediction of process and product quality, and (at present still semi-automatic) the parameterization and configuration of the technical systems for process and quality optimization are being forecast. Figure 22 shows a selection of Industrie 4.0 use cases taking advantage of AI technologies.

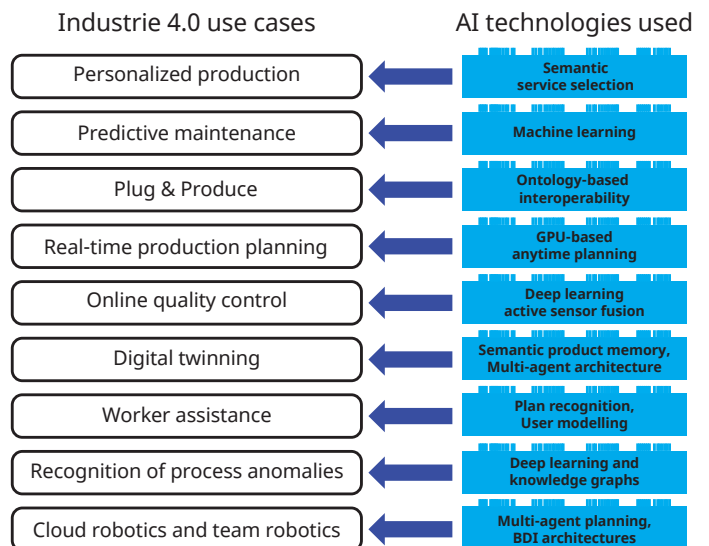


Figure 22: Examples of industrial AI used in selected Industrie 4.0 applications

The standardization of AI in Germany is of great importance – not least because of the national AI Strategy of the German government.

For this reason, the topic of AI has already been addressed explicitly and specifically in Version 4 of the DIN/DKE Standardization Roadmap Industrie 4.0 [217]. In standardization of AI in industrial applications, a distinction must be made between horizontal and vertical aspects. Horizontal standards are valid across different areas of application, e.g. generally applicable technical rules for measuring the quality of (technical or information) systems. In contrast, standards exist in different application areas (vertical standards), such as Industrie 4.0. In these areas of application, specific technical rules are developed which reflect the concrete applications and specific requirements of the area of application.

#### 4.5.1 Status quo

In German industry, the topic of AI and related topics have been of great importance for several years. Associations such as VDMA, ZVEI and Bitkom as well as VDI and VDE are dealing in various working groups with different aspects and various applications of AI. A variety of different descriptions of the application of AI in the form of application scenarios or use cases are considered. For various reasons (e.g. lack of a uniform description methodology, heterogeneous points of view and greatly differing levels of abstraction), exchangeability and comparability are not yet given, neither nationally nor internationally.

The use of AI in industrial applications can, depending on the application purpose and function of the AI, influence the fulfilment of requirements described in standards and other technical rules. For example, if AI technology is used to adapt the behaviour of automated functions, the influence of AI on the automated system must be considered in the conformity assessment. This applies in particular to industrial applications with functional and industrial safety requirements. Consequently, it is necessary to always check and ensure the fulfilment of normative framework conditions, especially considering the function and influence of AI. An objective assessment of the AI's sphere of influence is particularly necessary in this context.

Within DKE/AK 801.0.8, a VDE application rule VDE-AR-E 2842-61-1 “Specification and design of autonomous/cognitive systems” [218] has been developed, in which

terms and concepts for dealing with autonomous/cognitive systems are defined. A reference model for system and application architectures is being developed that considers the entire life cycle. Some approaches from the field of functional safety are transferred to this reference system. IEC SEG 10 is dealing with ethical aspects in autonomous applications and AI as an important approach to technology acceptance. In particular, socially relevant aspects are being considered and recommendations for the IEC Standardization Management Board (SMB) are being developed.

VDI (the Association of German Engineers) covers a broad spectrum of engineering sectors. In this context, AI is a cross-sectional topic, which is why numerous professional societies and departments are dealing with this issue from different perspectives. As part of the work of the VDI GMA Technical Committee 1.60, a status report on machine learning [219] has been prepared focusing on the requirements of users in terms of scientific research and knowledge transfer. The VDI/VDE/VDMA 2632 series “Industrial image processing” [220]–[223] describes the drawing up of specifications and the acceptance of classifying image processing systems. These guidelines are currently being revised, since image processing systems with artificial intelligence behave fundamentally differently from conventional methods.

The Task Force “Usage of new technologies” of IEC/TC 65/JWG 23 (together with ISO/TC 184/SC 1) is carrying out an evaluation of new technologies and their relevance for standardization in the field of smart manufacturing. Here AI in industrial applications is being regarded as a future technology. Work in this vertical AI area is mirrored at national level in Working Group DKE/AK 931.0.14.

At European level, the Focus Group on Artificial Intelligence was established in CEN/CENELEC in April 2019. It advises CEN and CENELEC on the development and dissemination of AI in Europe and focuses its work on responding to specific European needs, while generally globally relevant issues are resolved at a global level where possible. Among other things, the CEN/CENELEC Focus Group will take into account the guidelines on artificial intelligence for Europe of the High Level Expert Group on Artificial Intelligence set up by the European Commission [224]. The CEN/CENELEC Focus Group is developing a common vision for European AI standardization. Within CENELEC Technical Committee CLC/TC 65X, aspects of the use of AI in industrial automation are considered at European level.

The currently high level of interest in AI is leading to a multitude of different activities within different associations, institutions, consortia and societies regarding the application and standardization of AI in the industrial sector. The Standardization Council Industrie 4.0 established the Expert Council for Artificial Intelligence in Industrial Applications in order to avoid parallel additional work in the standardization of AI for industrial applications, to promote the exchange between these different activities and, ultimately, to develop a national opinion that is as harmonized as possible. The objective is the national coordination and harmonization of standardization activities to develop a consolidated picture of requirements and standardization needs in the context of AI in Industrie 4.0 of German industry, and coordination of suitable standardization activities. The Expert Council for Artificial Intelligence in Industrial Applications is the centre for discussions and the coordination of technical regulation in the field of artificial intelligence for industrial applications. Its tasks include the collection of use cases and the derivation of standardization requirements based on them, the development and specification of recommendations for action and their incorporation in various national and international standardization roadmaps currently being developed and those to be developed in the future, and the coordination of national and international standardization activities.

## 4.5.2 Requirements, challenges

The fields of action and the structure of this chapter are based on the thematic organization of the Plattform Industrie 4.0 and analyze its content results from a normative perspective: Basic requirements (see 4.5.2.1), application scenarios and use cases (see 4.5.2.2), secure, trustworthy AI systems (see 4.5.2.3), data modelling and semantics (see 4.5.2.4) and humans and AI (see 4.5.2.5).

### 4.5.2.1 Basic requirements and challenges from the point of view of Industrie 4.0

Today, it is not yet possible to make a clear statement as to when AI-specific standards/regulations take effect in a system or component. One possibility of classification is the definition of autonomy classes. At present, existing definitions of autonomy classes lack objective criteria for assignment and, with this, the possibility of evaluating when AI-specific technical rules are used with new systems and/or components. This need has already been identified by the AI project group

of the Plattform Industrie 4.0 and has been further substantiated within the work of the SC14.0 Expert Council for Artificial Intelligence in Industrial Applications.

In the development and operation of modern industrial applications in which new technologies, especially those from the AI environment, are used, aspects from other disciplines, such as law, ethics or economics, are taking on an increasingly important role alongside classical technological aspects. However, terms in different disciplines are partly provided with different meanings, whose specific interpretations are unknown in an interdisciplinary context, or at least difficult to grasp. This is especially true in the standardization of artificial intelligence in the context of Industrie 4.0, which is why technical rule-making for industrial automation requires a suitable and accepted common language and terminology (glossaries, ontologies). In this way, all stakeholders can be made familiar with the terms as comprehensively as possible, so that on the one hand a common interdisciplinary understanding emerges, and on the other hand, user safety is guaranteed in the cooperation between human and machine, which may be controlled by an embedded AI.

### 4.5.2.2 Requirements and challenges regarding the preparation and concretization of AI in industrial applications

In the Standardization Roadmap Industrie 4.0, structured preparation by means of use cases and scenarios was already recommended in order to prepare the scope of application of artificial intelligence in a structured way, to concretize the successful economic and technical use of AI in industrial applications, and to be able to derive standardization requirements tailored to applications. Following this recommendation, existing AI use case collections were examined with a focus on the manufacturing industry and a structuring or classification of these use case collections was undertaken. In particular, it became apparent that individual use cases could not clearly enough work out which innovations result from the use of AI compared to the use of classical models and methods, or why the problem underlying a use case cannot be solved without AI. This is partly due to an unclear understanding of the term AI, which can be addressed by scoping the term “AI” and by a clear classification of use cases (especially for the manufacturing industry) in order to clarify when it is an AI use case.

The analysis also included a categorization in order to distinguish use cases from application examples. The focus of collected examples is on the design and optimization of internal production value creation processes using artificial intelligence. The aspect of new business opportunities through AI has not been adequately addressed, so appropriate examples should be used to illustrate how AI can be used to create new business opportunities in industrial production.

In the AI environment, one often speaks of AI companies which offer cross-sector solutions as providers of technologies, algorithms and methods from the field of AI. Currently established companies in the manufacturing industry hire AI companies and continue to take the business risk and/or expand their portfolio through AI, but are not displaced by AI companies. To date, very few concrete examples are generally known in which an AI company has assumed business responsibility and the associated business risk in industrial applications. This aspect should be better illustrated by suitable examples and thus be examined more closely.

As already mentioned, the structured examination of AI use cases in the manufacturing industry serves, among other things, to identify possible standardization needs. However, most use cases are described very briefly and generically and therefore do not have sufficient depth of detail to allow the derivation of any requirements for standardization. Therefore, use cases should be described in sufficient detail to enable standardization requirements to be derived.

The collection of AI use cases carried out so far within the scope of the work of the SCI4.0 Expert Council for Artificial Intelligence in Industrial Applications has already made it possible to derive the recommendations for action described above. To systematically derive concrete recommendations for action with regard to relevant standards and specifications, the work already carried out should now be continued in a systematic consolidation process (with regard to the number, degree of coverage and quality of the use cases) so as to create a representative use case collection. Furthermore, the existing use cases should be detailed from a functional, technical perspective in order to identify relevant standardization relationships and to establish targeted cooperation with expert committees on the subject, which would ultimately allow concrete recommendations for action in technical regulation to be drawn up.

If one considers the use cases which have already been and are being developed within the framework of

IEC/TC 65/WG 23, these largely cover the value-added processes production planning/engineering, production execution and product service, which are predominant in the manufacturing industry, but the equally fundamental value-added processes product design and product configuration/sales are not covered. In order to achieve a complete coverage of the value-added processes in Industrie 4.0, it is necessary to complete the use cases already developed in the context of IEC/TC 65/WG 23 with regard to the topic of AI.

#### **4.5.2.3 Requirements and challenges regarding secure, trustworthy AI systems**

In the industrial context the proof of necessary properties (e.g. robustness, explainability, etc.) is of essential importance for IT security and safety. In AI, “black box” machine learning methods are often used, such as neural networks, which are very susceptible to small changes in the input data. Among other things, adversarial attacks make use of this. Established methods of verification, such as code reviews, are no longer possible through the use of black box technologies. The use of formal (mathematical) methods is a possible approach to gain knowledge about the internal relationships. Albeit there are projects in various committees which either focus on the use of AI without reference to industrial applications (project 24029-2 of ISO/IEC JTC 1/SC 42) or, as in the case of IEC/TC 65/SC 65A, IEC/TS 61508-3-2, which is currently being developed, takes into account industrial applications but does not consider the use of AI. Consequently, there is a need for action regarding the suitability testing of formal methods for the demonstration of specific properties for the use of AI in industrial applications. In concrete terms, a corresponding need for action was identified, for example, in the VDI/VDE/VDMA 2632 [220]–[223] series of guidelines on industrial image processing. Requirements and functional specifications must be created differently if systems with artificial neural networks are used. The same applies to the acceptance and testing of the classification performance of an image processing system.

Due to the increasing dynamization of value-added networks and the associated cooperation of systems during operation, the number of potential configuration variants is increasing massively and it is no longer possible to consider each individual configuration a priori. Digital twins or administration shells, which enable reconfiguration at runtime from a functional point of view, must therefore be enhanced with safety features so that the risk assessment of a configuration can

be performed at runtime. In general, the aim is to minimize worst-case assumptions about the system environment in order not to compromise performance unnecessarily. Possible solutions include conditional safety certificates (ConSerts) [225] and digital dependability identities [226]. A basic idea here is to replace worst-case assumptions, which apply in all situations, with situation-dependent assumptions that can be checked at runtime. Such approaches must be tested in industrial practice. In addition to the uncertainties in the system environment, the uncertainties in system behaviour pose a major challenge for reliability. The application of methods like machine learning leads to unpredictable system behaviour. Simple monitoring mechanisms which limit the system behaviour regarding reliability, are often not applicable because they are not situation-specific and limit performance in many situations. Therefore, monitoring mechanisms must be researched that can identify and control the risks of the current situation. AI methods can be used to automate safety-critical tasks that previously could only be performed by humans. Due to the complexity of these tasks, one cannot assume that the error rate is as low as for very simple safety functions such as an emergency stop switch. However, the number of accidents could still be significantly reduced if a task could be performed significantly safer by an AI than by humans. This raises the research question of whether existing risk acceptance criteria are suitable for AI-based safety functions or whether new concepts should be introduced to minimize the number of accidents.

The application of AI for (industrial) systems with safety functions currently poses a great challenge because corresponding standards and guidelines do not sufficiently consider the use of AI. For example, it is often misunderstood that the use of AI is prohibited from SIL2 upwards. Furthermore, IEC 61508 [79]–[86] states that safety is achieved as soon as the safety function has been implemented. The key question is what is right or wrong for an AI. There are no clear rules as to what an AI may and may not do and what evidence must be provided by an AI. This results in a clear lack of clarity for various stakeholders, such as users, solution providers, certifiers, etc. For this reason, a revision of relevant standards and guidelines in the field of safety and security, in particular IEC 61508 [79]–[86], is absolutely necessary for the use of AI.

The influence of AI on legal and regulatory frameworks is currently being discussed on a political level in Germany and the EU (see “White Paper on Artificial Intelligence – A European Concept for Excellence and Trust” [15] and the “Report on the Impact of Artificial Intelligence, the Internet of Things

and Robotics on Security and Liability” of the European Commission [227]). In particular, safety and liability issues are considered and specific aspects are discussed, such as specific regulatory requirements for high-risk AI applications, changes in the function of a product after it has been placed on the market by AI systems that learn in operation, specific requirements for human oversight over the entire life cycle of AI products and systems, and transparency regarding the development and behaviour of AI systems. This can result in interactions with AI standardization activities which have to be considered.

#### 4.5.2.4 Requirements and challenges regarding data modelling and semantics for AI systems in industrial applications

Today AI use cases are primarily described syntactically, i.e. the expressions have only a freely chosen (ontological) meaning, which makes it difficult to describe the dynamics of the use cases. Current, semantic models based on established vocabularies or such that have been newly developed in the course of projects are defined as predominantly static architectures of instances. Possible consequences of interactions between model instances (narrations) are often insufficiently described. Currently, interactions between models in terms of a targeted (re-)combination can only be formalized in highly individual and thus hardly transferable approaches. To address this challenge, a narrative representation based on a declarative semantic style should be used to provide a consistent description for the (re-)combination or compatibility of partial models.

Today the possible relationship between components is annotated at best case-by-case, for example, which portions of the interfaces can serve for the support of interactions. Any patterns according to which such interaction possibilities can be designed are also highly individual and allow little assurance of matching at the model level. It is also not possible to check in advance whether models from different sources can work together in the context of quality assurance, for example. For this reason, the use of narratives in the development process is recommended in order to indicate in models which elements can be changed, and to what extent.

#### 4.5.2.5 Requirements and challenges for the cooperation between humans and AI in industrial applications

As described in 4.5.2.1, there is a need for a common, interdisciplinary understanding of all, partly very heterogeneous aspects of the application of AI in the manufacturing industry. From this requirement for a uniform semantic view of industrial plants including data, processes, interactions between humans and machines, and the linguistic expression for ontological characterization, there has emerged a new need for a vocabulary with rules of application (guidelines) with which formal and calculable expressions or a language can be understood both by the machine and by humans in their own way. In this context, the principles of common logic are often mentioned, which are sometimes difficult for experts to understand, meaning they are unable to the application of formalisms. This has resulted in the need to describe the principles of common logic and its role in the standardization of AI for Industrie 4.0 in a way that is geared to application.

Another question requiring an answer is the influence of the use of AI on the work of engineers in the various disciplines. Since AI is or can be used for a variety of tasks, this question is becoming relevant in more and more technical areas. Accordingly, more and more working groups will deal with the influence of AI on the work of engineers in their respective fields. This challenge will be met by various committees – especially by associations and societies – in the context of the application of AI in an industrial environment. VDI has already announced its special commitment in this area.

#### 4.5.3 Standardization needs

##### Standardization and technical regulation

###### NEED 1:

###### **Criteria for the classification of systems or components within the framework of AI**

It is proposed to define criteria for differentiation (e.g. from existing automation systems). Existing divisions according to autonomy classes could be extended for this purpose.

###### NEED 2:

###### **Criteria for the classification of use cases considering the role of AI**

Clear criteria are needed when it is an AI use case and when not. A clear argumentation is necessary as to why certain needs for action have been identified precisely because of AI.

###### NEED 3:

###### **Adaptation of existing standards, specifications and guidelines**

There should be an evaluation as to whether, and to what extent, existing standards, specifications or technical rules need to be modified to align them with AI. As examples of such activities, the clarification of IEC 61508 [79]–[86], the revision of Machinery Directive 2006/42/EC [94] and the final draft of ISO/CD TR 22100-5 can be considered.

###### NEED 4:

###### **Standardization of a data standard for economic interoperability of declarative models**

It is proposed to explicitly mark the life cycle and interaction relevant elements and their interrelationships (patterns) in models to support uniform automated processing. This is similar to OWL-S in that it includes not only descriptions of interfaces but also their associated structures. Standards are needed for these actions and their handling.

###### NEED 5:

###### **Standardization of a formal I4.0 methodology that supports the principles of declaration and narration in combination**

A standardized procedure/method is to be defined as to how the actions named in individual models under Recommendation for action 4 can be specifically taken, evaluated and assessed before an upcoming application context. This should make it possible to carry out a priori plausibility studies of subsequent potential interactions.

###### NEED 6:

###### **Standardization of a design process for models to be described with semantic formats such as declaration and narration**

It is proposed to define a development process that supports the design of models for the purpose of later dynamic interconnection (i.e. it is known at the time of model creation that the model should interact with other models, but not yet how).

**NEED 7:****Regulation and liability**

The influence of AI-specific adaptations of regulation and liability law on AI standardization activities should be considered.

**NEED 8:****Review of the legal and regulatory framework for safety-critical tasks**

A suitable adaptation of legal and regulatory framework conditions reinforces the use of AI technologies and enables, among others, medium-sized companies with calculable economic risks to use this technology.

## Research

**NEED 9:****Risk assessment by AI/New methods for risk assessment/ Dynamic risk management**

There should be an examination and evaluation of the extent to which current methods of risk assessment and requirements are not yet adequately addressed by the present standards, specifications and technical rules. As an example of this activity, the work on ISO/CD TR 22100-5 can be considered. This examines risk assessment methodology according to ISO 12100 [137] (hence also Machinery Directive 2006/42/EC [94]) from the aspect of the potential impact of AI on safety.

**NEED 10:****Collection of terms from different disciplines (glossary)**

It is proposed to develop a common language (i.e. semantics) in the form of a glossary with rules, laws and axioms that provide clear definitions both for specific disciplines (e.g. law, technology, economics) and across industries.

## Application

**NEED 11:****Expansion of the collection of use cases regarding new business opportunities through AI**

Future examples should include not only the design and optimization of internal production value-added processes, but also examples in which new business opportunities are opened up on the basis of AI.

**NEED 12:****Identification of business scenarios for the role of AI companies in industrial automation**

It is proposed to specifically identify and prepare business scenarios in which an AI firm takes a business risk in relation to the value proposition of AI.

**NEED 13:****Standardized preparation of use cases**

It is proposed to prepare use cases according to the IEC/TC 65/WG 23 template and to achieve a necessary level of detail of about 20 pages per use case by preparing the use cases according to the usage viewpoints of the IIRA template.

**NEED 14:****Considering specific use cases and the role of AI for product design and configuration**

It is proposed to develop further use cases, for example for the value-added processes product design and product configuration/sales, which are currently not being addressed by IEC/TC 65/WG 23.

**NEED 15:****Checking the coverage of collected use cases with the observation scope of IEC/TC 65/WG 23**

Before developing further use cases (e.g. informative machine, adaptive logistics), it is suggested to first check to what extent they are already addressed by existing use case descriptions (e.g. IEC/TC 65/WG 23).

**NEED 16:****Updating the collection of use cases in a national coordination committee**

It is recommended to continue adding to the use case collection created in the course of the work of the SCI4.0 Expert Council AI in Industrial Applications.

**NEED 17:****Detailing existing use cases**

Further detailing of use cases according to usage view in the form of functional views.

**NEED 18:****Formal methods**

Testing of the suitability of mathematical methods for proving necessary properties (e.g. robustness, explainability, etc.) of black box machine learning models.



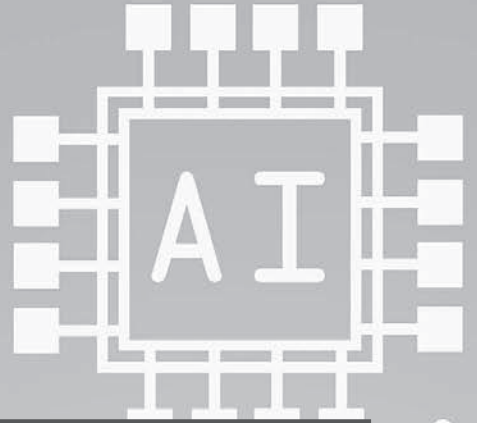
**NEED 19:****Application of plasticity and elasticity of models with regard to conflict detection/criticality**

It is proposed to apply an explicit distinction in the development process of models, to identify which elements can be changed, and to what extent, in order to interact with other models. These distinctions shall be used to make predictions about the expected fit of two models at runtime of a system.

**NEED 20:****Evaluation of the principles of common logic and its role in the standardization of AI in I4.0**

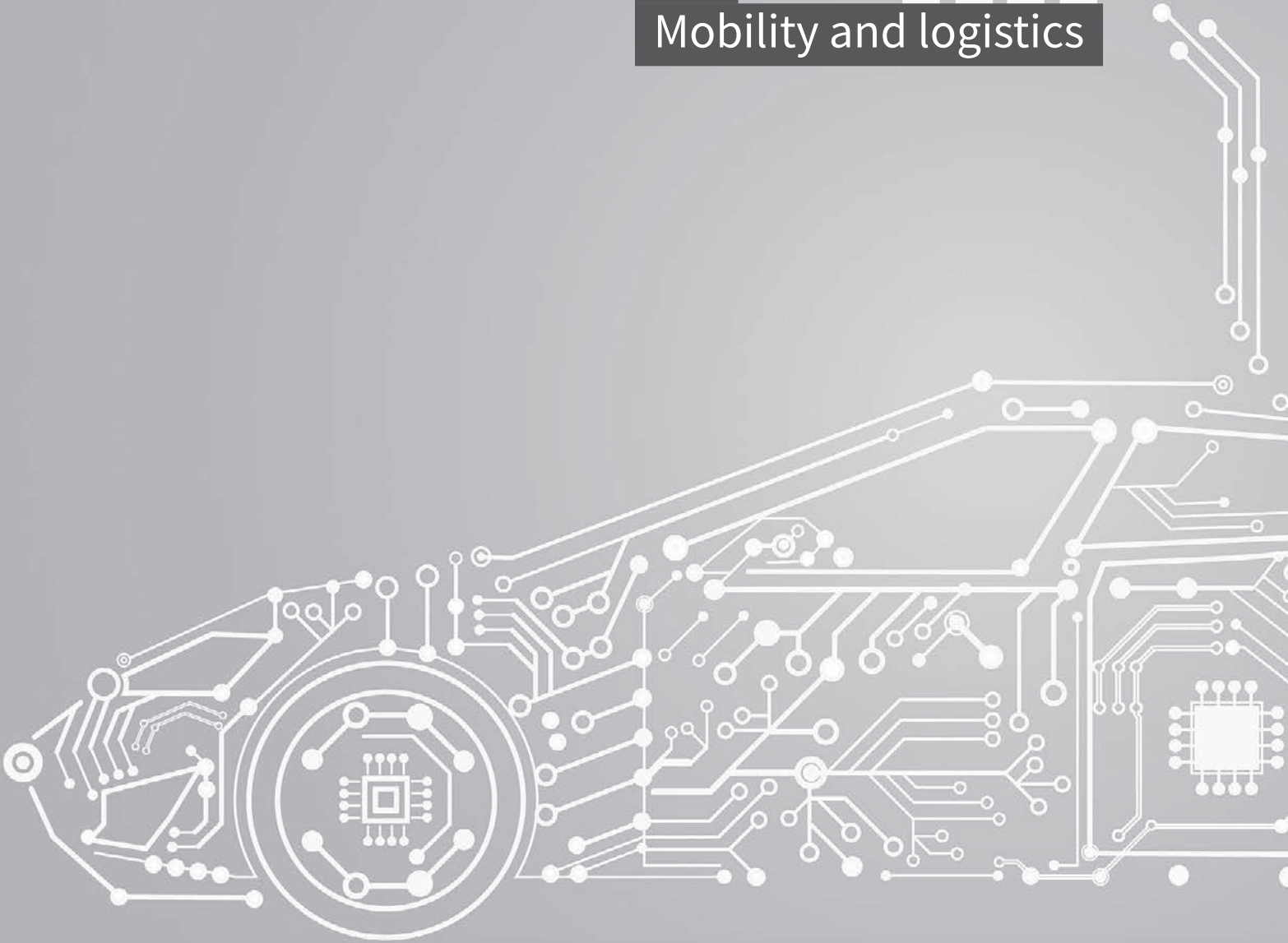
Description and outlining of the application of the principles of common logic/semantics and its role in standardization. Strengthening of networking with all participants, comparability with alternative approaches and community building to bring these activities to the broad mass and general spectrum of I4.0.





**4.6**

**Mobility and logistics**



This chapter explains the massive innovation potential and some of the associated rapid changes that the use of AI brings to the domain of “mobility and logistics”. The chapter is structured along three essential aspects:

**Legal framework:** Relevant aspects of the existing legal framework for mobility and logistics are briefly outlined. This serves as a basis for demonstrating the potential of standardization. It is recommended that all parties involved in the process (industry, testing organizations, legislators and approval authorities) work together to examine AI-specific applications in the light of the prevailing legal framework.

**Explainability and validation:** As in other domains, explainability and validation of AI systems play a major role in mobility and logistics. A large number of standardization and research committees are currently dealing with the question of the depth to which the AI algorithms used must be documented and explained in a comprehensible manner so that a clear functional relationship between input and output can be recognized. Linked to this is the question of proving correct functionality (validation). Due to the complexity of the issues, the increased demand of applications and the requirement of standardization to reflect the current state of the art in science and technology, the task is to define further fields of research in order to prepare standardization topics.

**Interoperability:** Making IT systems interoperable is common practice, not only in the domain of “mobility and logistics”. Nevertheless, the topic of interoperability deserves special attention in the discussion of AI systems for mobility and logistics. In this domain there are heterogeneous and multimodal services and applications based on systems of different operators and providers, which will grow massively in their functionality through AI. For this reason, standardized data models and interfaces will make a particular contribution to innovation and efficiency.

#### 4.6.1 Status quo

**Road Traffic Law:** In the Federal Republic of Germany, the Road Traffic Act (StVG) [228] was extended in 2017 to include automated driving (which still requires the presence of a driver to take over), i.e. for high and full automation. In addition to the obligations directed at the driver and the definition of high and full automation, the new StVG contains references to certification law, because only vehicles with certified high or fully automated driving systems fall under the amended Road

Traffic Act. In particular, the law requires that the highly or fully automated driving system complies with traffic regulations. This requirement has recently been included in the technical regulation of Automated Lane Keeping Systems (ALKS) and is included in the scope of type testing. If AI mechanisms would now change the driving behaviour of the highly or fully automated driving functions in terms of the German Road Traffic Act, the type approval already granted would no longer cover the new driving behaviour. If one wanted to change this mechanism in the future, this would have far-reaching (legal) consequences.

At international level the Technical Committee ISO/TC 22 “Road vehicles” is already working on ISO/TR 4804, *Road vehicles – Safety and security for automated driving systems – Design, verification and validation methods*. AI is also being addressed in other ISO/TC 22 standards. The scope of this work is under discussion.

In addition to the activities in standardization, the consortial standard UL 4600 “Standard for Safety for the Evaluation of Autonomous Products” [157] is currently being developed. The non-profit organization Underwriters’ Laboratories (UL) is responsible for this.

Furthermore, work in the field of automated driving has started in a working group (IEEE P2846).

A cooperation between the standardization fields is based primarily on the topic of logistics. A joint working group (cooperation between CEN and CENELEC) has been applied for at the European standardization level with the participation of DIN and DKE. This initiative is gaining significant relevance as the ratio of transport routes (currently approx. 70 % of all goods worldwide are still transported by road) is changing and cooperations are becoming more dynamic and complex.

In relation to this, China is already involved in the European standardization organization CEN as a cooperation partner. Both partners intend to develop common standards for the transfer of European goods.

Specifically for the food sector, international work is underway on the forthcoming standard ISO 23412, “Indirect Temperature Controlled Refrigerated Delivery Services – Land transport of parcels with intermediate transfer”, under the leadership of Japan, which considers the transport of frozen goods (mainly fish and beef). The intended standard can be

seen as a further cornerstone for the automatic distribution of goods worldwide.

#### 4.6.2 Requirements, challenges

The challenges for AI systems in mobility and logistics are complex and profound. Examples are automated driving, control of international flows of goods, optimization of warehouse logistics, and scheduling of rail vehicles.

##### 4.6.2.1 Legal framework

Homologation/technical regulation in the field of motor vehicles is as follows: In Europe, motor vehicles that are allowed to participate in public road traffic without restrictions must meet the requirements of a type approval. This is regulated by the Framework Directive 2007/46/EC (from 1.9.2020 in the EU Regulation 2018/858 [229]) and the UNECE Agreement of 20 March 1958 [230] on the adoption of uniform conditions for the approval of equipment and parts of motor vehicles and on the reciprocal recognition of approvals. This Agreement, which comprises over 150 technical regulations, was a milestone on the way to achieving uniform technical approval regulations. These concern not only systems and components for active and passive safety but also environmentally relevant regulations for the protection of all road users. The type approval means that the equipment or parts have been proven to fulfil all necessary regulations according to the EU type approval regulation and that they have been tested and confirmed by an independent testing institution (in Germany TÜV, DEKRA etc.) appointed by a national type approval authority. Parameters that are not relevant for type testing or do not affect it are excluded. An approval authority issues the type approval on the basis of this confirmation, without which the vehicles may not be placed on the market. The manufacturer confirms this with the Certificate of Conformity (CoC). Type-approved vehicles with a valid certificate of conformity are registered as road users in Germany or in other EU member states and are issued with a license plate. Periodic technical inspection in accordance with EU Directive 2014/45/EU [231] ensures that the vehicles meet all applicable safety and environmental requirements even after many years of use. It is therefore an important means of creating consumer confidence.

However, in the course of technical development and digitalization, products or their function can even be changed during

their useful life. The currently valid processes for approval, type approval and periodic monitoring must therefore ensure in future that even these changes do not impair road safety and that environmental protection is still guaranteed. New and efficient mechanisms must be developed with which changes to the product during the use phase can be recognized, tested and released. The certification processes will become much more product-specific in the future. One challenge is to adapt type approval and periodic technical monitoring to automated and networked vehicles.

For the approval of semi-automated vehicles, the regulation for steering systems (UN-R79 [232]) has been comprehensively revised and expanded since 2016. In a first step, requirements for systems for assisted lane keeping and lane changing, for corrective steering interventions in case of imminent lane departure or to avoid objects, and for automated or remote-controlled vehicle parking were added. The systems described belong at most to the category of partial automation (automation level two)<sup>29</sup>. In the meantime, it has been decided that the requirements for highly or fully automated lane guidance systems (automation levels three and four) should not be issued as a supplement to UN-R79, but as a separate regulation for ALKS.

The ALKS regulation formulates important boundary conditions:

- Limited to highways
- Limited to traffic jams and low speeds of 0-60 km/h
- Limited to lane guidance systems, no automatic lane change possible
- Limited to vehicles of registration category M1 (passenger cars), application in the commercial vehicle sector is planned.

In order to resolve these far-reaching limitations, the UNECE working group “Functional Requirements for Automated Vehicles” (FRAV) has been working since October 2019 on generic requirements for the approval of automated driving functions. In parallel, the “Validation Methods for Automated Driving” (VMAD) working group has been working since March 2018 on standardized methods to demonstrate compliance with these requirements. According to the current state of discussion, this proof should consist of an audit and a simulation part, as well as additional physical tests on test sites and an assessment drive on public roads.

29 [https://www.sae.org/standards/content/j3016\\_201806/](https://www.sae.org/standards/content/j3016_201806/)

To describe possible validation scenarios, a uniform format was developed in the German funded project PEGASUS in the form of openSCENARIO and openDRIVE, which is now available via ASAM e.V. as a freely accessible standard and which has been incorporated into the work of the UN. This section of the text was essentially consolidated by the working group “AG 6 – Standardization, Specification, Certification and Type Approval” of the National Platform Future of Mobility, NPM AG 6 (see 3.2), and has since been published.

The aspect of machine learning (ML) is particularly relevant with regard to the type approval and periodic technical monitoring (main inspection) of AI systems in motor vehicles, procedures that are regulated by European law. It is expedient to distinguish AI systems according to whether this learning takes place during development or during operation of the systems. Based on the ongoing ISO/IEC CD 22989 project, the term “trained model” should be used to refer to “learned systems” or “offline learning” in cases where machine learning takes place exclusively in the development phase and the system does not change after it is placed on the market. By analogy, “continuously learning systems” or “online learning” should be spoken of when machine learning takes place during operation, i.e. when the system changes after being placed on the market.

Special attention is paid here to the training data of AI systems. There is a need for regulations or standards on the provision and use of data collections (voluntary data exchange or data pools), which can be used for training purposes of AI systems.

Product liability and product safety law responsibility, as well as the responsibility for compliance with type approval regulations and market access rules (summarized as “product compliance”) lies with the manufacturer of the product in which the AI system is used or “installed”. According to § 9(2) of the German Product Safety Act (ProdSG) [195] a product which complies with standards or other technical specifications is presumed to meet the requirements of product safety if they are covered by the relevant standards or other technical specifications. This paragraph applies analogously to road and rail vehicles.

#### 4.6.2.2 Explainability and validation

The ability to validate the behaviour of AI systems is a prerequisite for their integration into safety-relevant products. Vali-

dation must be carried out in accordance with the state of the art. The state of the art for the validation of AI systems is still developing dynamically, so that no corresponding standards exist. Thus, the state of the art is only implicitly defined. This makes it difficult to decide when validation is sufficient.

Furthermore, the further development of the mobile phone standard enables the transmission of large amounts of data from the vehicle. This development enables a new and timely tool for the continuous validation of mobile automated systems in the context of field monitoring. It is important to find out how these new possibilities can be used in a target-oriented way. In particular, continuous validation after market launch could help to detect deviations from target behaviour and identify potential for optimization.

Explainability supports validation. Transparency and traceability of decisions made by the AI system and of the decision-making process the AI system goes through. In order to be able to understand decisions of AI systems, both aspects of the development (e.g. data and training methods used) and the execution (e.g. crucial characteristics) of the AI system must be considered. Benchmarks and metrics for the traceability and transparency of AI systems should be developed.

So far, there are no clear definitions for necessary interfaces of an AI system to the human being or society and to (another) AI system(s). The increasing complexity and breadth of automated functions makes it difficult for a human observer to distinguish early on between proper functioning on the one hand, and deviations that require intervention on the other. Since the decision of an automated system can be based on heterogeneous sensor data, internal model calculations and networked data exchange, for example, it is not immediately traceable to humans per se. Furthermore, a person’s relationship to the system influences their ability to judge and their considered possibilities of intervention. This ranges from the operator of the system (e.g. the owner of an automated car) to the casual user (e.g. the passenger of an automated subway) to the involuntarily affected person (e.g. a pedestrian versus an automated vehicle). Thus, a system potentially interacts with a heterogeneous set of affected people with different expectations of the system. Standardization of human-machine interfaces can help people to understand the inner state of a concrete system independently of their experience with it, and to assess whether they need to take action themselves and which actions or interventions are possible, necessary and/or appropriate.

We have now arrived back at the topics of transparency and traceability. When an AI system gives the decision to a human being, the human being must know how, when and from which decisions the transfer is being made.

#### 4.6.2.3 Interoperability

A multitude of business processes in mobility and logistics (e.g. intermodal transport, third party logistics services (3PL services), public transport or traffic flow control) depend on the close cooperation of the respective actors across organizations. Where the necessary systems and processes are not (or should not be) directly linked, interoperability becomes a key factor in successful business relationships. The goal of interoperability is to make the cooperation between actors as efficient and effective as possible in order to minimize frictional losses during interactions (e.g. time delays, queries, misunderstandings, format conversions). The more interoperable heterogeneous systems are, the lower the interaction effort and error rate and the greater the flexibility and resilience of the overall system – characteristics that are becoming increasingly important in times of advancing digitalization, the emerging Industrie 4.0, upcoming structural changes in logistics, big data and artificial intelligence. This is all the more true since collaboration – for which interoperability is a key success factor – is becoming increasingly critical to success in today’s highly dynamic environment. Future standardization activities, especially in the context of artificial intelligence, will therefore consider interoperability as a principle and clearly define corresponding standards in order to minimize scope for interpretation and thus increase the compatibility of business processes.

Interfaces have a special importance for the interoperability of AI systems. The focus is on intermodal transport chains for people and goods, as well as the planning of such transport chains, e.g. goods from warehouses to public roads, recommendations and implementation for individual movement of people (These examples are described in greater detail in the application scenario on intelligently networked mobility of the Platform Learning Systems [233]). As soon as AI systems cooperate with each other, there must be clear rules on what this cooperation looks like, since each AI system is also independent. It can be assumed, for example, that it is necessary to think beyond the interfaces between the AI systems and, for example, to name the objective functions of participating systems in order to exclude unwanted interactions. Another major challenge is seen in the data. Among other things, it is

still open how high its level of quality is, how it is collected, with which data an AI system initially learns (training data), and how the technically and legally simple data distribution via cooperation chains should look for usability by all.

Data exchange between individual systems forms the basis of automated processes, making interoperability a key success factor for Industrie 4.0 and digitalization. In the context of artificial intelligence, interoperability primarily refers to software, software interfaces and data as well as their compatibility with each other. Due to the speed of development in these areas and the multitude of different development strategies, (new) software solutions and interfaces often engender new challenges to practical implementation or assurance of interoperability. For standardization, this means that standards must always take up existing best practice approaches and/or widespread solutions. In addition, system-independent design principles and “rules of the game” for interoperability should be found that are – as far as possible – timeless.

The particular challenge in using artificial intelligence is above all **to identify the data that has been generated by AI-supported applications**. Co-operation partners must be able to distinguish whether the transmitted data results directly from real data, for example from an ERP system, or from a calculation of an AI system. It must also be possible to assess the context in which the data was created and how reliable it is (analogous to the “quality of service” principle). This is especially true when current and future systems have a high degree of autonomy, i.e. act largely independently: The lower the level of human involvement in the control and monitoring of these systems, the higher the requirements for interoperability and especially for the quality of service of the AI-supported generated data.

In all standardization projects, it should be noted that interoperability is fundamentally a **holistic challenge**. It should be considered and taken into account from the very beginning – from design and test procedures to implementation and daily operation. The more this is achieved, the less effort is required to subsequently improve the interoperability of different systems, facilitating the realization of economic, ecological, social and safety-related potentials.

All activities to ensure interoperability as well as its implementation in practice (e.g. data collection, storage and exchange) must comply with the applicable **data protection** framework.

### 4.6.3 Standardization needs

#### 4.6.3.1 Legal framework

##### NEED 1:

##### Implement solidarity of the process participants

It is expedient to involve industry, testing organizations, legislators and approval authorities in equal measure in order to discuss AI applications from the field of mobility and logistics under the aspects of the prevailing legal requirements (vehicle regulations, legal requirements – e.g. from the StVO [234], transport of goods and merchandise – as well as relevant regulations such as the GDPR [95] in the area of data exchange, continuous vehicle inspection). The objective should be an interpretation of the legal framework for future AI applications and, if applicable, a strategy for adapting the legal framework to enable novel AI applications in society.

##### NEED 2:

##### Need for clarification on “safe and compliant” for provision on the market

All AI systems used in the field of mobility and logistics must be “safe and compliant” before being made available on the market, as well as during their useful life (“useful life” in this context means the life cycle of the vehicle or the life cycle of the function/service). Here there is a distinction between products and services. Products must comply with the relevant laws and regulations. Whether these laws and rules are decisive for a pure AI system service (“software service”) requires further clarification.

##### NEED 3:

##### Resolve contradiction between “static” starting point and “dynamic” learning

The use of self-learning AI systems at execution time would create considerable challenges under current law, because the (functional) properties could change in a way that is difficult to trace and predict due to the self-learning. The treatment of such constellations under product liability law is still under discussion. It would also be difficult to cover such products in terms of certification or regulation law, as the law has always required/demanded a fixed point of reference for product characteristics.

Before installing software updates on systems already in use, it is important to analyze their current status under the aspect of possible changes in order to minimize unintended safety-relevant interactions. In addition, changes to motor vehicles, for example, can also become relevant for approval,

which could require prior testing by those involved in the process. It is necessary that process participants discuss the system-related legal/regulatory situation for those cases in which the use of AI systems generates changes at runtime.

#### 4.6.3.2 Explainability and validation

##### NEED 4:

##### Promote research

When testing AI systems in the area of mobility and logistics, clear definitions of the test criteria, the test process, the test identity and the exact test contents forming the basis from which the test is developed are required.

For the preparation of corresponding standardization, the following research tasks and their promotion are recommended in this respect:

- Research into the risk that systems to be tested are specifically optimized for tests, e.g. that AI systems are trained for singular situations and over-adapt to test contents (“learning by heart”, “over-fitting”)
- Development of tests including dynamic test procedures which counteract the above-mentioned risk of optimization
- Characterization of AI systems that are self-changing through learning in use and/or are used in changing environments; corresponding impact on continuous testing

##### NEED 5:

##### Accompany research

Research projects for unambiguous, generally valid and objective evaluation criteria and methods under consideration of a continuous validation of the safety and performance of automated and networked driving with increasing use of AI must be actively supported and accompanied. Suitable algorithms for the evaluation of the driving task should be developed and their interfaces defined. AI methods can especially be used. The results – e.g. from the PEGASUS project, legislative working groups (e.g. IWG FRAV and VMAD of the UNECE) – are to be incorporated into standardization and are to be based on human driving behaviour.

##### NEED 6:

##### Transparent design of AI systems

To make AI systems transparent, specifications for the execution time are needed. Relevant points are those that evaluate the usefulness of interactions with other systems, as well as the competence of the AI system for the current situation.



For the preparation of standardization in this field, the following research tasks are recommended:

- Methods for identifying and describing the AI system’s own area of competence (e.g. adversarial examples, context and limits), especially for safety-related functions or transition to a safe state.
- Comprehensive analysis of the human-AI interaction (e.g. AI suggests different options, human selects another, or conflict area “safety vs. security”) in a certain action, i.e. the traceability of the action of the AI system.
- Research on how neural networks can be used for safety-related functions. This concerns their development and release process, as well as detection methods for properties and explainability. Furthermore, the question arises as to which architectural patterns are appropriate for the integration of neural networks into safety-related functions.

For the automotive sector, the “AI assurance” project from the VDA’s lead initiative Autonomous and Networked Driving addresses questions regarding the protection and release of AI systems for a specific application. Similar initiatives are recommended for other application areas.

#### 4.6.3.3 Interoperability

##### NEED 7:

##### Create data reference model for interoperability

Data and its correct use is a crucial success criterion for interoperability. In logistics and mobility, there is already a wide range of best practices for data, data types, data models and databases (e.g. master data, change data and transaction data, their relationships to one another and possibilities for integration in software solutions). These best practices are currently changing due to new technologies, requirements, possibilities and solutions. Standardization committees and users should observe daily practice and, if a new best practice, e.g. of data types, becomes apparent, define it uniformly, taking into account a certain short- and medium-term flexibility, in order to propose a data reference model for interoperability in mobility and logistics: In such a model, basic data types relevant for interoperability, their structures and relationships to each other should be described, taking into account certain degrees of freedom (e.g. by “should” or “can” provisions). Existing work, for example on metadata (e.g. ISO/IEC 11179 Metadata Registry [235]–[242]), should be taken up. A data reference model would make it possible to develop use cases as well as interfaces with greater speed

and compatibility and also provide a uniform (communication) basis on which the multitude of actors involved in mobility and logistics can standardize the data aspect of interoperability. This data reference model in combination with a function reference model (see below) can be a reliable basis for interoperability.

##### NEED 8:

##### Create a function reference model for interoperability

Due to the necessarily holistic view of interoperability and the simultaneous diversity of standardization organizations and initiatives, coupled with the great social and economic importance of mobility and logistics (not least as a critical infrastructure), a functional reference model for interoperability should be designed and anchored in standards as soon as possible in order to create a uniform understanding of what distinguishes interoperability in the context of AI applications and how it can be realized and ensured. Furthermore, the functions required to achieve interoperability, such as data acquisition, processing, evaluation, transfer, etc., should be defined and their basic systemic requirements explained. In addition, proposals should be developed on how to ensure that these interoperability functions are implemented in accordance with requirements from the design phase of a system throughout its entire life cycle. This functional reference model should integrate the data reference model interoperability (see above) and also take up existing work (e.g. ISO/IEC 19763 on the metamodel framework for interoperability (MFI) [243]–[252]) and proven methods and tools (e.g. systems modeling language (SysML) or unified modeling language (UML) for modeling). The continued suitability of these models is to be evaluated under the condition of further development and integration of AI solutions in daily practice. The findings and results of this process should also be incorporated into non-AI-related work on interoperability in national and international standardization organizations.

##### NEED 9:

##### Specify methods for data exchange

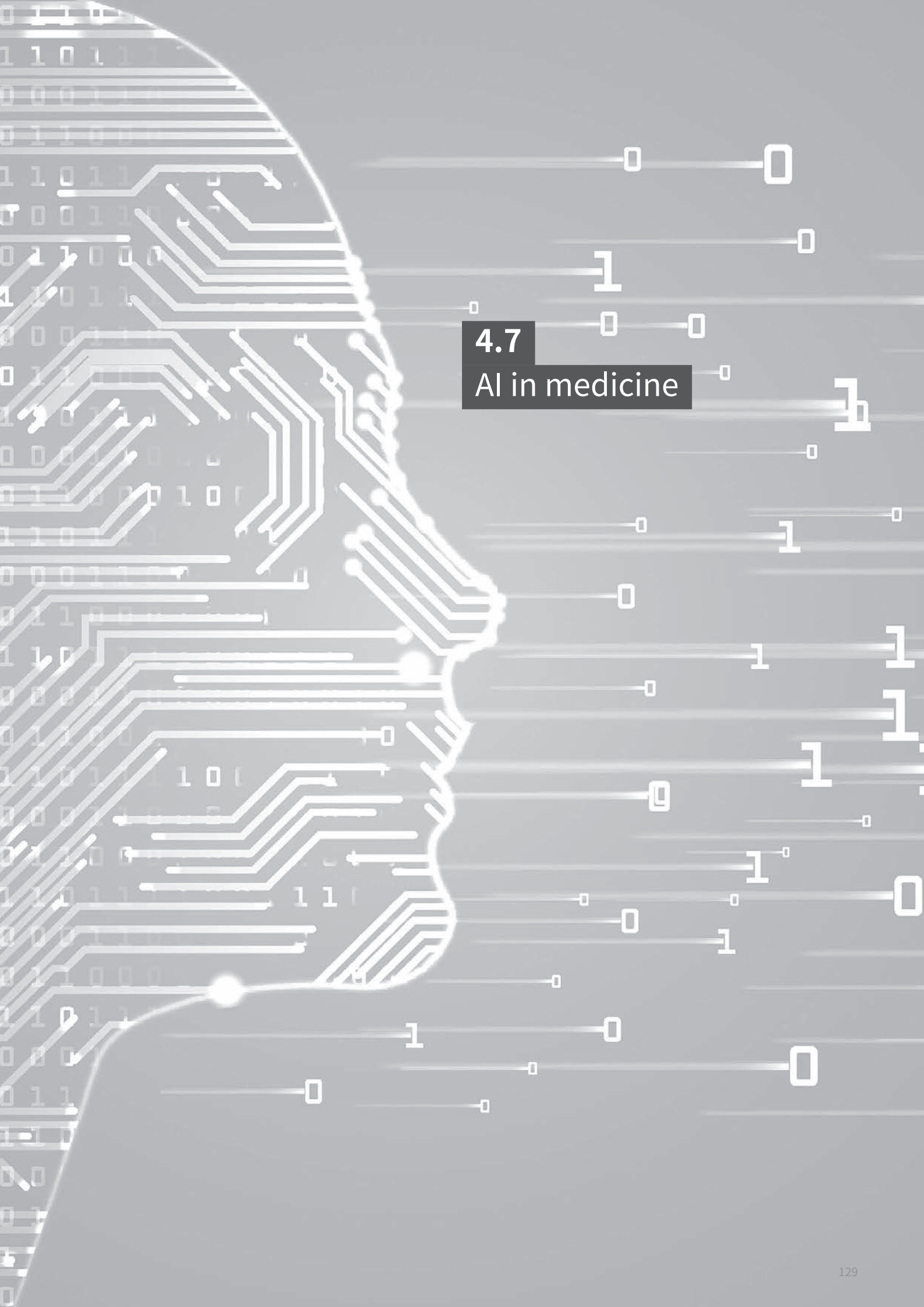
The interfaces are the supporting point for interoperability. Against the background of increasing data volumes and data complexity, as well as the progressive use of data by AI-supported tools, data exchange procedures should also be standardized, especially with regard to syntax, semantics, formats, consistency, coherence, completeness (e.g. quality of service information on AI-generated data or quality level) and type of data transfer, so that actors can optimize their systems and interfaces accordingly. Furthermore, the exchange of supplementary data should be standardized in order to be

able to transfer, for example, data models, inference engines or information on the autonomy of the integrated systems between actors as required, thus allowing interested users to independently verify data quality etc. In addition, quality seals and methods for their award (including the necessary quality criteria, test mechanisms, etc.) should be defined in order to create the possibility of building interoperability on a solid basis of trust. This would reduce the need for independent audits and thus have a positive effect on actor relationships and sustainability.

### **NEED 10:**

#### **Define the type and quality of data**

Due to the expected increase in networking between actors and the increased use of (partially) autonomous systems, the minimum type and quality of data required to ensure interoperability should also be standardized. This includes clear guidelines on how and to what extent given data (types) are to be enriched with e.g. further information on the data context, what limits apply to the autonomous further processing or use of these data, or from what quality level onwards data are suitable for autonomous processing. This also includes the definition of quality criteria (quality of service) which data must meet in terms of interoperability and security. Corresponding specifications can be included in the data reference model.



4.7

AI in medicine

In medicine, AI is creating further options for prevention, diagnostics and therapy: from smart apps for the – currently – early detection of diseases to even more differentiated, personalized oncological therapies. In order to take advantage of such opportunities, suitable secure framework conditions must be created. In addition, there are still challenges to be mastered in the field of ethics, legal context, economy, technical aspects, acceptance and empathy. Is it ethical to listen “to a machine” in sensitive questions about modalities of survival and life and death? What regulations should there be so that technology always serves people – and not the other way around?

#### 4.7.1 Status quo

For the field of medicine, the standardization-related preliminary work and results have so far been clear; whereas the technologies for medicine and health are already extremely complex on the market. For example, health apps offer advice on medical questions from private individuals and professional users worldwide. There are also chatbots in medicine for the analysis of diseases.

Current systems fall into the area of “weak AI” and are developed in the fields of “knowledge-based systems”, “pattern analysis and pattern recognition” and “robotics”. (Classification according to: [12] pp. 4 ff.).

With different classification systems for imaging procedures, diagnoses – e.g. in the fields of laboratory diagnostics, parasitology, radiology, pathology, cytology, dermatology, ophthalmology – can be made faster and more precise, or paths in minimally invasive surgery, for example, can be made safer. For some surgical procedures, surgeons already have the possibility of using robotic systems, also cooperating via telemedicine for specific indications.

In hospitals, clinics, doctor’s offices and institutes, AI systems are not only used in diagnostics or therapy. AI-based systems can also provide support in other areas. These include application areas such as exo-skeletons and prostheses, sensor-based monitoring and therapy monitoring, as well as solutions that improve processes in medicine and/or administration and thus help to make patient care more efficient and, in some areas, even possible at all (e.g. “intelligent” prosthesis systems).

The World Health Organization (WHO) and the International Telecommunication Union (ITU) are cooperating in a focus group called AI4Health and have already published a white paper [253].

Deployment of AI in medicine is dependent on a number of prerequisites. The action needed may be to:

1. Unify classifications, standards and terminologies for healthcare data;
2. Ensure interoperability of data and check information and possibilities of data acquisition (e.g. mobile devices of patients, technical devices in hospitals, clinics, medical practices, health care institutions);
3. Establish cooperation among the different actors to establish binding standards promotes and clarifies data access, data origin, interests/demands on data, and interfaces;
4. Clarify open questions on infrastructure aspects with regard to “independent” medical infrastructure or “general” infrastructure with a specific module for medical data, whereby the constitutional and federalist context (federal state level, national, European, international) also appears to require clarification.

#### 4.7.2 Requirements, challenges

AI systems in medicine are significantly influenced by three factors:

1. the availability and quality of health data;
2. the legal framework; and
3. medical ethics.

All three imply the trustworthiness of an AI system if there is clarity about how privacy and transparency are ensured.

“Secure AI systems for medicine” is the title of a white paper from the Platform Learning Systems which recommends, among other things, the certification of AI systems for the secure use of AI systems in medicine and for the benefit of patients – for example, to ensure unaltered training data [254]. It seems important here to develop common guidelines and test specifications for the approval and certification of AI databases and their operators. In addition, manufacturers should be legally obliged to remedy defects and neutral institutions should be commissioned to operate the AI assistance system; all in all, a highly complex and challenging job. An independent test committee (e.g. notified bodies) can also check the functionality of the certified and deployed AI

systems at regular intervals. Recall processes could also be established.

In medicine, doctors and members of other medical and health care professions act according to ethical principles. For some time now, discussions have been taking place on this basis as to the extent to which it is permissible for devices to significantly influence or even take over decisions. Ultimately, doctors make diagnoses and determine therapies. Furthermore, there is a need for a debate as to when “human action” and when only “human oversight” (weak AI) is necessary, or when autonomous action is required (strong AI). There are simply no principles and regulations for human-machine interaction in the medical sector.

The ethical aspects are all the more complex the more globally an AI system is to operate. Different approaches exist worldwide due to different cultural and historical backgrounds and medical care structures. For Europe, the “Guidelines for Trustworthy AI” of the European Commission and, for Germany, the report of the Data Ethics Commission should be mentioned here in particular. With the Declaration of Helsinki (1964, last updated 2008 [255]), the World Medical Association (WMA) has defined a denominator for all cultures worldwide.

As in other areas, the question of liability also arises for medicine. This is the case, for example, when making a false diagnosis or causing personal or economic damage. This also results in uncertainties about the reversal of the burden of proof. In the potential – not to be desired – case of damage, there is a need of clarification whether customer, manufacturer, operator/user or a completely different involved agent is liable to prove. This requires a risk assessment, e.g. in scenarios (see DIN SPEC 92001-1 [87]):

**Table 9:** Requirements and challenges of AI systems in medicine

AI Module class	High risk	Low risk
Mandatory	No deviation from requirements allowed	No deviation from requirements allowed
Highly recommended	Deviation from requirements with justification only	Deviation from requirements with justification only
Recommended	Deviation from requirements with justification only	Deviation from requirements without justification allowed

On the one hand, approval is affected by ambiguity in the legal framework. In Germany there is a highly regulated Medical Devices Act (MPG) [256], in Europe the Medical Devices Regulation [141] with detailed regulations. Until now, any medical device may only be made available on the market if it has an approval, which in turn only considers a specific state. For the approval itself, medical devices must meet all relevant legal requirements and have undergone a conformity assessment procedure, possibly involving a notified body. The conformity assessment procedure refers to a specific technical state of the product with corresponding functions. With the continuous learning of AI systems and thus the modification of the product itself, the state at the time of approval (and certification by a notified body, if applicable) is sometimes already left behind.

On the other hand, the legal framework for the use of AI in medicine is unclear, for example with regard to civil liability for potential, undesirable treatment errors. Inseparably connected with this is the need for clarification of a permissibility of the use of AI systems for decision-making and also the renouncement of decision support by AI systems. Against this background, a broad social consensus is needed on approval and the use of continuous learning systems.

Thirdly, market access is characterized by a lack of clarity in the legal framework. Medical devices are sometimes very complex and strictly regulated. In Germany, the Medical Devices Act is currently still valid, but will soon be replaced by the EU Medical Devices Regulation.

As is well known, every medical device may only be made available on the market with a – nota bene valid – CE marking.

For CE marking, medical devices must meet all relevant legal requirements and have undergone a conformity assessment procedure, possibly involving a notified body. The conformity assessment procedure refers to a specific technical state of the product with corresponding functions. With the continuous learning of AI systems and thus the modification of the product itself, the state at the time of CE marking (and, if applicable, certification by a notified body) is sometimes left behind, which means that the conditions for market access are no longer met de jure.

Decisions of AI systems are based on the interpretation of existing data with necessarily high quality; however, this often does not seem to be given adequately at present. AI-based

applications are naturally subject to the GDPR [95]; this raises questions for the development and use of AI systems. Also in the current context, data ownership, data integrity and consent to data use are in need of clarification. Persons who consent to the use of their data usually do so for a specific purpose. By processing data in an AI system, this purpose can change inherently in the system. This results in the need for further clarification regarding legally compliant, ethically secured procedures, since an AI system can change with every learning step and cannot “forget” interpretations made with “unreleased” data, unless this has been provided for in the AI programming. The use of this data, e.g. as part of a learned model, has the potential to generate social benefits, e.g. in terms of improving diagnostic and therapeutic options beyond the individual, with further relief for direct and indirect participants in the health care system.

Data must be representative, consistent and accurate. In addition, technical availability is required (data formats, machine readability, security and access options).

### 4.7.3 Standardization needs

#### NEED 1:

##### **Define error classifications, misclassifications and learning from errors**

As a result of considerations of medical ethics, insights can be expected into which level in the AI system is essential for learning. Decisive for this are output data (input – with distinction in training data and full data, see DIN SPEC 13266), the actual learning strategy (processing) in the narrower sense, the result (outcome) and the use (impact). This requires a concretization for the preventive handling of misclassifications by an AI system, since an AI system can develop prejudices/ethically counterproductive processes based on its data and previous learning outcomes, and thus act in a misguided and/or discriminatory manner. The adequate handling of failed attempts, or whether learning through failed attempts is allowed at all, is also relevant for social-ethically compliant learning.

#### NEED 2:

##### **Define medical ethical values**

AI systems in medicine must adhere to ethical values in a socio-cultural context, whereby the specifications for this must be defined – if this has not already been done. An orientation is possible on the principles and ethical recommendations of the World Medical Association (WMA), which defines a com-

mon denominator for all cultures. These are contained in the current version of the Declaration of Helsinki [255] and are constantly being adapted. It must be possible for the AI system to comply with this declaration in the currently valid version. This demonstrates, among other things, the relevance of the connection between big data and ethical aspects.

#### NEED 3:

##### **Create a review process to evaluate existing principles**

An evaluation basis for the adequacy of given principles seems urgently necessary. As is well known, despite standardization by the World Medical Association (WMA), there are still different national principles, research results in different populations, different findings, etc. For an AI system it is necessary to develop a testing process that evaluates the suitability of principles, research results, findings, other variables, etc. and integrates the dynamics of AI systems (especially self-learning systems). This applies not only to the AI system itself, but also to the data quality.

#### NEED 4:

##### **Clearly define legal definitions and requirements for self-learning and self-developing/changing AI systems**

It must be clarified whether and, if so, how, proof (documentation) of the security and performance of such AI systems over the entire life cycle should be provided by the manufacturer. According to the current view of the interest group of notified bodies for medical devices in Germany (IG-NB), certification of independently learning and independently developing/changing AI systems is not possible under the current legal framework. Here the legislator is required to take action.

In the context of medical decisions and interventions on humans, such “continuous learning systems” develop, not least of all, a hitherto not clearly assigned, but on any account different ethical dimension. Legal definitions of responsibility are necessary – as well as definitions of liability for an AI system that develops/changes independently – focusing on manufacturers or, users or “third parties”.

#### NEED 5:

##### **Ownership, allocation and revocation of data**

The legislature is also called upon to regulate the ownership of (health) data and its allocation and, above all, the legitimate procedure for revoking or withdrawing a data transfer, integrating it into the data-based learning of the AI system that has already taken place.

It should be noted that, from a medical point of view, it can be valuable and value-adding to use an AI system in more applications than originally intended, or to use innovative solutions from another jurisdiction (e.g. use in Asia although data are from European institutions). Separate requirements for anonymization or pseudonymization are conceivable for the utilization of data. The second is another option when the original purpose is changed.

For institutions that want to make data available for research purposes, procedural, legal and technical questions arise in data extraction, quality assurance and data provision, which today can at best be identified by large institutions, but only partially answered

Furthermore, given the federal principle in Germany, the state data protection laws and the confessional data protection laws are a particular challenge in the interest of “only” national legal security; in any case, the European perspective (e.g. Health Data Space, Gaia-X, etc.) must be examined in particular.

#### **NEED 6:**

##### **Define data and its usage**

The data itself also requires further specifications, which must be made transparent to the responsible parties. For example, the manufacturers of data-based applications must define inclusion and exclusion criteria, provide a description of training, validation and test data and demonstrate how statistical “outlier data” are handled in a solution-oriented and legally compliant manner. The same applies to further additions, e.g. in connection with restrictions on working with analytical data – whereby it must be clarified whether restrictions are necessary or should be avoided. Relevant for this is a data set consideration of the effects of random variables with the goal of formulating definitions for individual data, statistical evaluations, etc.

#### **NEED 7:**

##### **Specify restrictions for big data**

The restriction of big data is to be clarified. Big data are initially the same as traditional data, except that it is generated in gigantic quantities. However, this then has an impact on the assessment of transparency and accuracy. However, while accuracy is increased due to the larger statistical sample, transparency decreases due to the variety of data and lack of reproducibility of the database used.

#### **NEED 8:**

##### **Balance data protection and data quality**

At present, it is advisable to increasingly include data protection because of the legal framework. Quality can suffer as a result. Thus, defined guidelines are needed to achieve a balance between data protection and data quality. In addition to the legitimate aspects of data protection law, however, the ethical imperative for the use of data must also be emphasized at this point, insofar as this serves the public good.

Data should be fair and must not discriminate. But there are still no guidelines on how this can be complied with in data collection. Data itself cannot assume a characteristic such as “fair”, but only the data set based on its collection (e.g. selection of the population in statistical surveys, evaluation of the person(s) evaluating the data, etc.). Defined guidelines reduce the risk that an AI system evaluates or acts in a discriminatory manner based on its data.

#### **NEED 9:**

##### **Generation and consensus of principles for human-machine-human interaction in the medical sector**

One orientation option is offered by the criteria for the design of the human-machine interaction [257], especially focusing the criteria on the protection of the 257 and trustworthiness. The focus is on data security, data protection and non-discrimination, as well as quality of available data, structured transparency, explainability and consistency of AI systems.

#### **NEED 10:**

##### **Promote innovations for the use of AI systems**

AI systems and applications not only provide enormous opportunities for better or consolidatable health care, they also offer sustainable prospects for Germany as a business location. A cultural change in the sense of a maximum promotion of innovations for the use of “artificial intelligence – Made in Germany” – also in health care – is a sine qua non in the interest of an internationally leading role; in this context, cooperations between science and industry are to be promoted which commit themselves to structured transparency and consensual openness in the development of AI systems and applications, and which network start-ups, small and medium-sized businesses, as well as large companies.

#### **NEED 11:**

##### **AI standards and AI excellence clusters for processing medical data**

Germany has one of the best health care systems in the world. The high level of technical equipment for imaging

procedures for diagnosis and therapy plays an important role in this respect. Despite large investments in modern X-ray equipment, ultrasound equipment, magnetic resonance imaging and other nuclear medical imaging techniques, only a very small percentage of the images generated daily are used for AI research and training of medical AI classifiers, aggravated by the lack of standards for processing medical images. Furthermore, the current German regulations on data and patient protection do not allow the use of images, text or voice documentation and therefore do not allow competitive research. All these factors mean that the potential to develop AI-based assistants holistically for the health care system, e.g. to provide information on prevention and to support doctors in diagnosis and therapy, is dwindling.

Therefore, a legal initiative is needed that enables and consciously promotes the use of medical data for regionally typical, community-based research and development purposes in health care. Furthermore, defined AI standards and AI specifications are necessary (see [Chapter 4.7](#)). In addition, an AI excellence cluster for medical imaging procedures could help to develop these standards and specifications. First projects should start in 2021 to reduce the gap to other countries.

#### **NEED 12:**

##### **Make data available to research**

In addition, there is a need for improved access to high-quality data for universities, colleges and other scientific and research institutions, as well as for companies that contribute to the future-oriented design of the health care system by developing innovations – for diagnostics, therapy and the pharmaceutical industry alike. In this context, corresponding activities at national and European level (so far the German Medical Informatics Initiative, the European Health Data Space) have to be defined, which especially include the ethical framework conditions for AI and to which future measures should be oriented.

#### **NEED 13:**

##### **Automated text and speech recognition procedures for approvals**

The regulations for the approval of drugs, medical devices and various measures are a complex construct of development, testing, conformity assessment and certification. A great deal of documentation effort is needed in this connection, slowing down the approval process despite urgent needs – especially in times of crisis. Here, an AI system that recognizes text or speech fully automatically can significantly accelerate the process.

#### **Epilogue/Summary and Outlook**

The outlined topics, recommendations for action and suggestions are not limited to diagnostics and therapy, though these are central areas. The focus is also on prevention, health promotion, “care” (an English term that covers more than the German term “Pflege”), screening, and multi-professional client/patient orientation.

An important aspect is the balance in (medical) ethics between “high tech” and “high touch”.

The great potential for benefit that AI offers in the medical and other fields is vitally dependent on reaching consensus across the whole of society – and a standardization roadmap is the safest possible way to achieve this together.



## 5

# Requirements for the development and use of standards and specifications

## 5.1 Review and development of standards and specifications in AI

### 5.1.1 Review of existing standards and specifications

The fields of application for AI are extremely diverse. In almost all economic sectors and also in other fields of application, AI technologies are used both in the form of components in end products and services and in the productive core and support processes within companies. According to the German government, artificial intelligence will thus sooner or later be of great importance for all economic and social areas. The situation is similar with standards and specifications. These also exist for almost all economic sectors and fields of application. The German body of standards currently comprises more than 30,000 standards (DIN, DIN EN, DIN EN ISO/IEC). Combining both theses means that a large part of the more than 30,000 existing standards must be reviewed and supplemented with AI aspects.

The Federal Government's AI strategy [12] addresses this aspect in Field of Action 12 and recommends the review of existing standards and specifications for AI suitability. Although the term "AI suitability" is not defined, it is meant thus: Standards and specifications that are relevant for the application of AI have to be identified and ultimately supplemented with AI specifics. By extending the scope of the standards and specifications, AI solutions can be used safely and reliably with their help. To implement this measure, a methodology is needed to identify existing standards and specifications with regard to their AI relevance. It may be possible to use IT tools to support this research. At the same time, a systematic approach must be developed to identify any need for action to optimize existing standards and specifications. Finally, on the basis of this preparatory work, measures are to be designed that aim at a comprehensive inclusion of AI aspects. One of the biggest challenges in this context is probably the lack of AI expertise. The often vertically oriented committee structures in which standards are developed, especially for traditional industries, require in-depth domain knowledge. This has to be extended by AI technology knowledge. It should be noted that the relevant standards in question are predominantly of European or international origin. In principle, the review of existing specifications is likely to be more difficult, since only a few specifications have been developed by the established standards organizations and the vast majority of the relevant specifications have been developed by various consortia.

### 5.1.2 Agile development of standards and specifications for AI

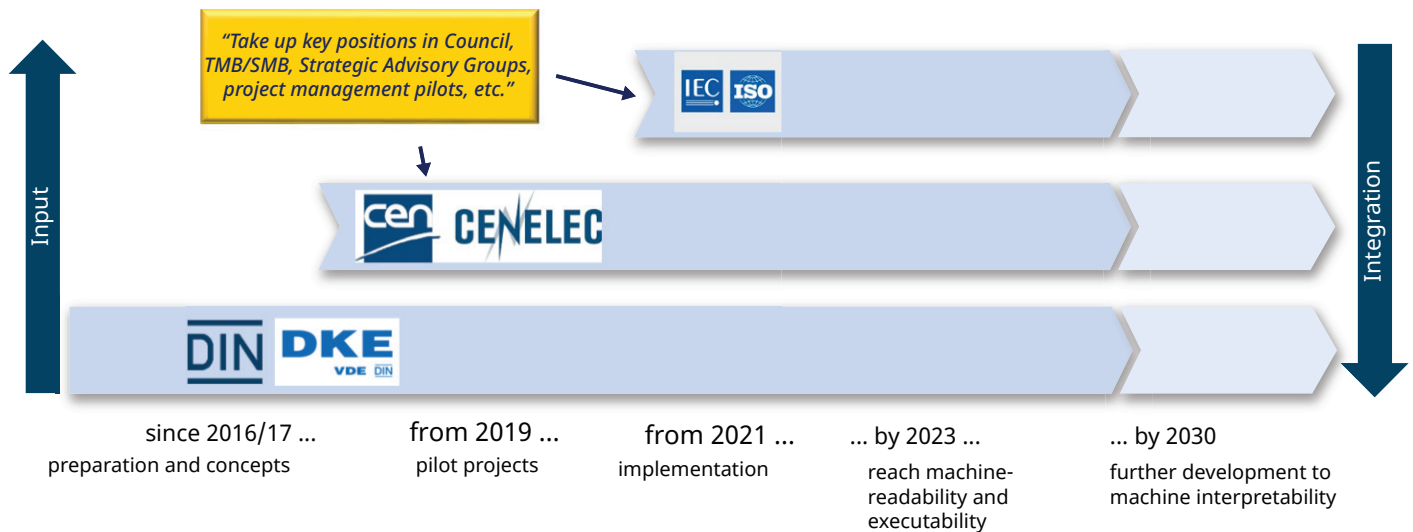
A major challenge for the development of standards and specifications for AI systems has been the great dynamics of AI technology development. Many industries use different AI technologies depending on the field of application of the AI solution and relating to the use case. Hybrid AI solutions are often even based on a combination of AI methods. In most cases, the specifics of the application are met by state-of-the-art approaches from AI sub-disciplines, which are individually adapted and refined. Consequently, the dynamics at the interface between AI research and industrial development and application are particularly high. In this way, the applied AI is constantly being developed and industrially evaluated. AI standardization must take this tension between applied research and industrially mature development into account and pursue pragmatic, bidirectional approaches in the analysis of standardization needs and the development of market-ready specifications. This requires an iterative process which, in the design of standards and specifications, incorporates reciprocal impulses from research, industry, society and regulation and supports continuous and mutual learning between the actors. At the centre of this approach is the testing and successive refinement of the developed specifications along use cases. In this way, application-specific requirements can be identified at an early stage and marketable AI specifications can be realized. As a result, the acceptance of AI specifications by industry, science and society is ensured.

## 5.2 SMART standards – New design of standards for AI application processes

This section presents the motivation for SMART standards in relation to AI, as well as the current state of development of a future model and possible technological approaches to the realization of SMART standards. In addition, [Annex 11.4](#) contains a more detailed description of the technological approaches for further reading.

### 5.2.1 Motivation

**Definition of SMART Standard: Standard, the contents of which are applicable to machines, software or other automated systems (Applicable) and readable (Readable)**



**Figure 23:** Standards organizations involved in SMART standards

and, in addition, can be provided digitally in an application/user-specific manner (Transferable).

SMART Standards – this is a topic that has been gaining importance for three years, not only nationally in DIN/DKE, but also in CEN/CENELEC (CCMC) and ISO/IEC, see [Figure 23](#).

By means of coordinated pilot projects and scientific studies, the promoters of this innovative task and its possible solutions are taking initial steps. One challenge will be the consolidation of a common target picture of standards producers and users: What does a future development process of SMART standards look like and which information model is required? What input can SMART standards provide for downstream AI-based application processes? This contribution summarizes the existing knowledge on these issues and provides approaches and impulses for further investigations, which are marked as P requirements with AI relevance in the text and annex.

### 5.2.2 Status quo

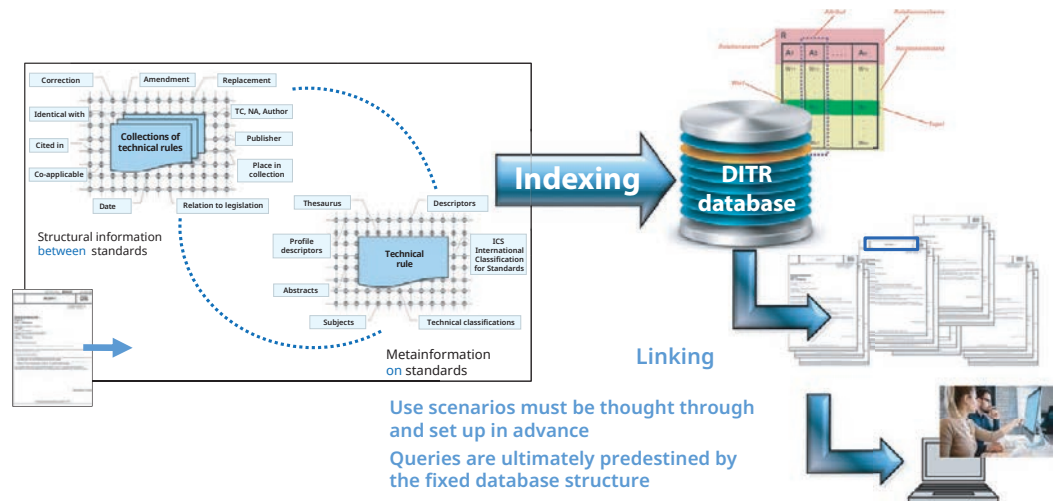
The direct further use of standards and their contents in downstream processes is gaining increasing attention. Companies expect efficiency gains in the future [258] from standard components (value tables, part descriptions, 3D models, software, requirement definitions, test procedures), which can be directly taken over and executed by machines. Comprehensive electronic provision of national/European/international standards still takes place today mainly via

PDF-based standards management procedures organized by means of metadata. From a technological and user point of view, we are at a mature and reliable level – for a meanwhile very broadly based provision of information with regard to the number of listed collections of technical rules and regulations and the depth of indexing (see [Figure 24](#)).

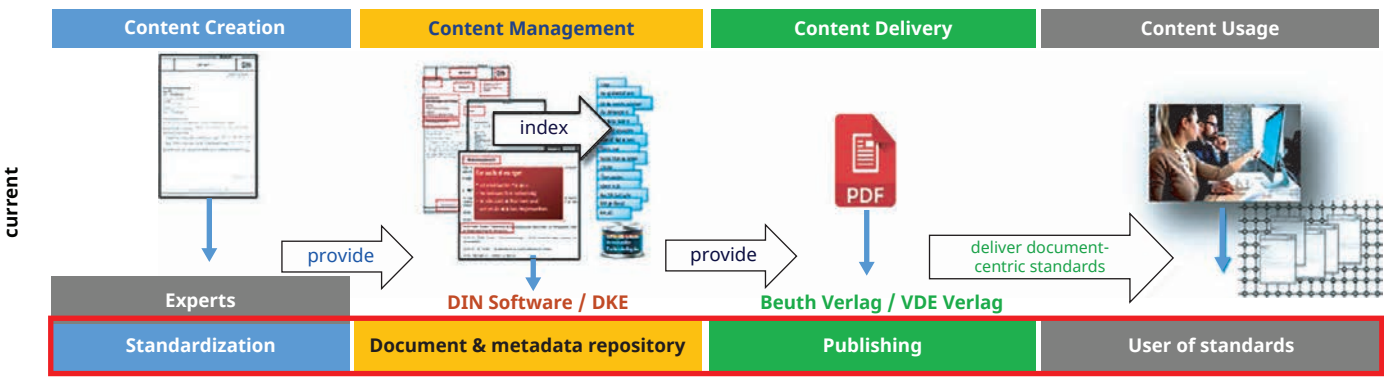
Today’s workflow, which has been established for decades, functions successfully and in a balanced manner on the basis of silent or explicit agreements between the process partners involved. The underlying principles are carefully coordinated in accordance with standards and legal requirements and guarantee reliable management of standardization results in customer-oriented systems. Today’s changes of the values of the parameters “principle” and “characteristic value” are carried out conscientiously and by consensus of all parties involved, taking into account the applicable rules, see [Figure 25](#).

Today’s workflow, which has been established for decades, functions successfully and in a balanced manner on the basis of silent or explicit agreements between the process partners involved. The underlying principles are carefully coordinated in accordance with standards and legal requirements and guarantee reliable management of standardization results in customer-oriented systems. Today’s changes of the values of the parameters “principle” and “characteristic value” are carried out conscientiously and by consensus of all parties involved, taking into account the applicable rules, see [Figure 25](#).

**Figure 24:** Partially automated indexing of documents for customer processes



**LEVEL 1**



**Some relevant requirements for the existing overall process**

**The “standardization principle”**

- A stable process described in detail, which has successfully asserted itself, e.g. as regards rights to content, the participation of stakeholders, publication of drafts, public commenting, etc.
- Standardization is based on principles, e.g. consensus, uniformity, internationality ...
- Based on quality characteristics, e.g. type of legally binding nature, anti-trust principles, consumer acceptance, democratic legitimation, product liability ...

**Key figures (in DIN e.V.):**

35,000 experts from industry, research and the public sector work together with 200 project leaders in DIN on 2,000 new standards per year (of a total of 34,000 German Standards)

**The “metadata principle”**

- Centralized operation and maintenance of the metadata database
- Process (indexing) based on standardized rules agreed for decades
- Professional exchange of experience with key customers established

**Key figures (e.g. DIN Software):**

Continually developed, highly automatic process with ca 90 metadata fields in 300 collections of rules (national and international) leads to ca 60,000 changes to records per year which are administered by 20 specialists

**The “service principle”**

- Dissemination of national and international standards and other technical rules
- Development or provision of expert knowledge in all media formats for industry, science, commerce, services, study and the trades ...
- Offering services to support the customer’s processes

**Key figures (e.g. Beuth Verlag):**

800,000 products (incl. national and international standards) which are offered by 180 staff members to 170,000 active customers

**The “user principle”**

- Use of structured and reliable data for managing documents (drawn up acc. to the standardization principle) in customer-oriented systems
- In addition to the publicly available standards, companies can also draw up internal standards
- The use of standards is voluntary and at the user’s own discretion

**Key figures (on the German market):**

Companies use ca 10 to 20,000 standards and solutions

**Figure 25:** Today’s workflow: From standardization to usage of standards

The upcoming far-reaching process-related changes in the context of the SMART standards development of content management, distribution and usage will have to be delimited and redefined against the background of existing introduced and regulated procedures. The decisive value (“asset”) of a standardization subject must be preserved.

The model has three dimensions:

- Presentation forms Level 1 to 4: “Utility model” – usage formats based on the required technologies, see also [Figure 26](#).
- Creation and usage scenario: “Process model” [260] – from content creation to content usage.
- Examples for the realization of the sub-processes.

### 5.2.3 SMART standards – Level model

One challenge will be to consolidate a common understanding among developers and users of SMART standards: How will a future development process of SMART standards be designed, what content structure is required and what are the application scenarios? On the basis of a model (see [Figure 26](#)), a systematic description of the sub-processes and their partial solutions is given below. The models are also presented in two detailed web-based seminars [259].

The existing value creation process for **Level 1** is already described in [Chapter 5.2](#). It is shown in [Figure 26](#) for completeness, in order to better classify the transformation of the individual subprocesses on Levels 2 to 4.

The activities at **Level 2 and 3** are first steps towards SMART standards, as granular information will be created within existing standardization processes. Further information is given in the Annex in [Annex 11.4.1](#).

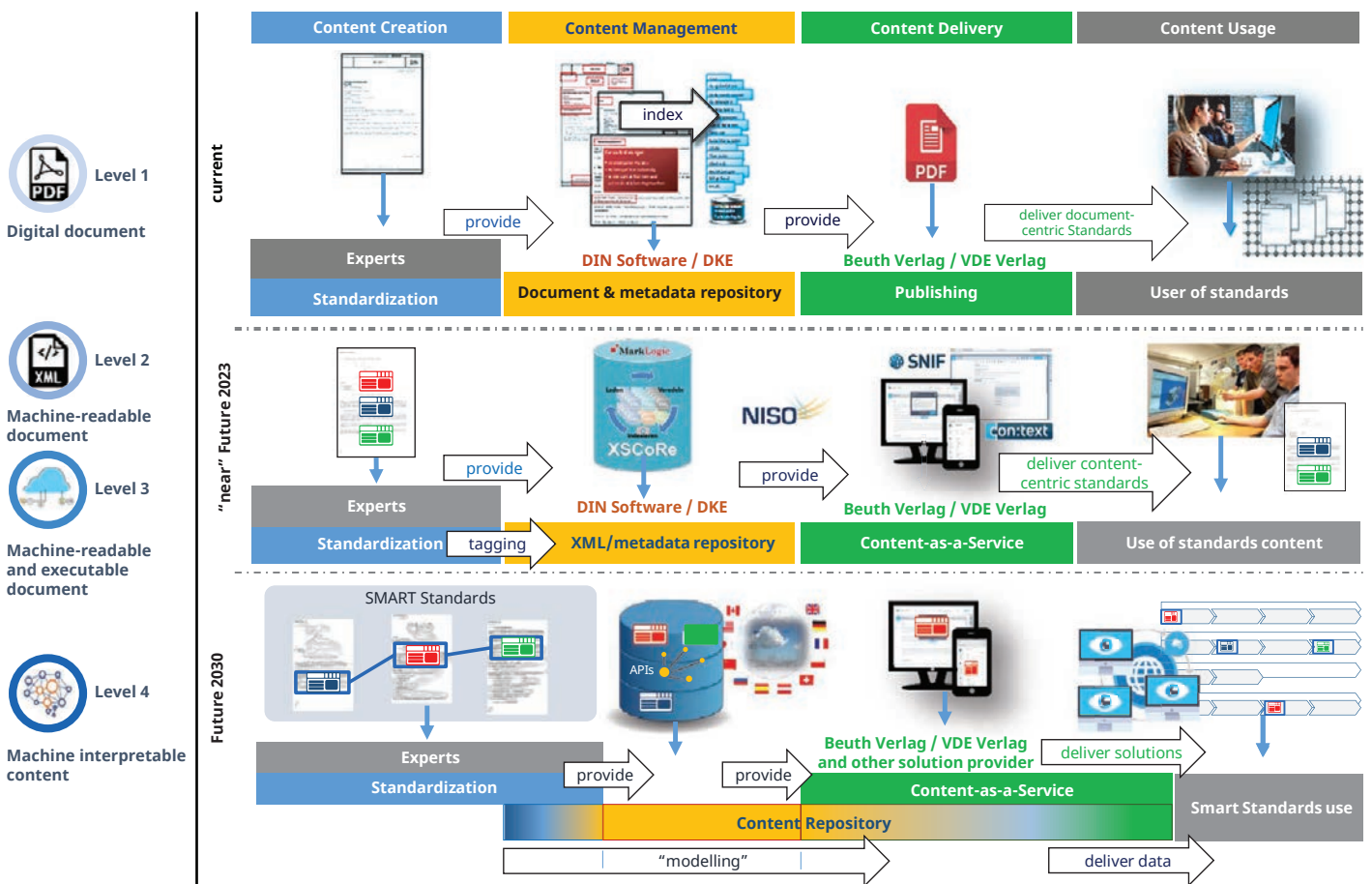


Figure 26: SMART standards level model

**Level 4** represents the final stage of a continuous SMART Standards value chain, see [Annex 11.4.3](#).

The level model must be verifiable and adaptable to other models, e.g. Reference Architecture Model Industrie 4.0 (RAMI4.0) [261].

Creating an open and constructive culture of discussion when introducing new processes also means: It is necessary to involve the users (users of standards) and the standardization community at an early stage and on an ongoing basis in the development of the methodology. In fact, the SMART standards project is supported and promoted “at all levels” at national, European and international level. Nationally the following activities around SMART standards have been or are being developed, e.g:

- **IDiS** (Initiative Digital Standards of DIN/DKE): SMART Standards of the Future (DIN): <https://din.one/site/sof>  
IDiS (DKE): <https://www.dke.de/de/normen-standards/digitalisierung-normung-digitalstrategie-dke-transformation>
- Cooperation with the German Committee of Standards Users (**ANP**): <https://www.din.de/de/service-fuer-anwender/anp>
- Cooperation with the **BFA** (User Specialist Committee of DIN Software GmbH): <https://www.dinsoftware.de/de/normen-management/benutzerfachausschuss>
- Work within the **NAGLN** (DIN Standards Committee Principles of Standardization): <https://www.din.de/de/mitwirken/normenausschuesse/nagln>
- Cooperation with various universities

In the above-mentioned committees, an important new aspect is being repeatedly discussed: How do you prepare those involved for the new requirements?

A further aspect of systemic relevance for the future concerns the definition of the requirements for the changed qualifications of the external but also the DIN internal “actors” in the overall process. Existing concepts must be further developed in order to describe the newly arising tasks of all process participants in SMART standard processes.

### 5.2.4 Standards and AI

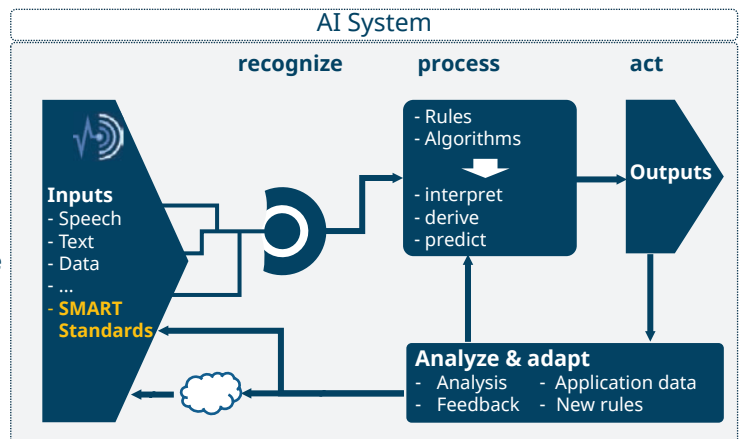
One of the goals of this project is the derivation of rules for formalizing and modelling the content of standards and specifications. The resulting improvements in the quality of the underlying data are essential for the optimal functionality of AI systems. The intended development of a central repository for structured standards data can serve as a basis for high-quality AI applications, see [Figure 27](#).

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.

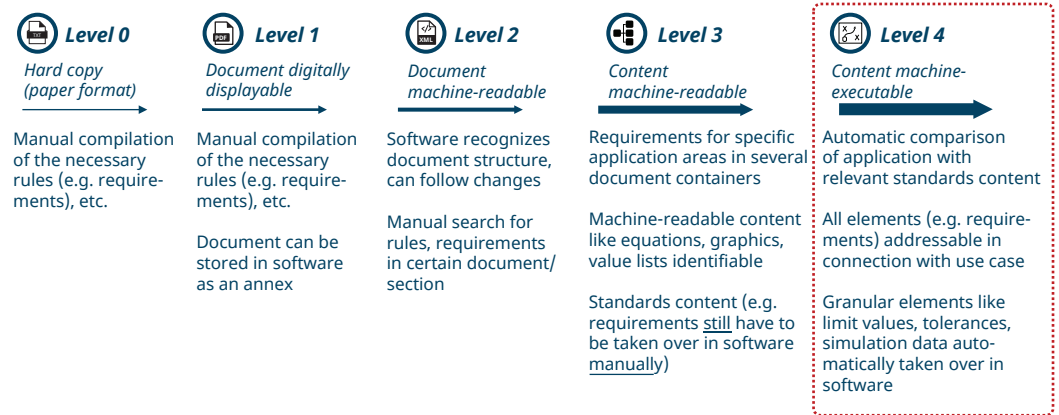
The conception of the necessary data models and interfaces will have to be part of this project and thus makes an important contribution to the further penetration of AI applications in the sub-processes of enterprises.

**Figure 27:**  
SMART standards as input for AI

- Numerous input formats for learning/self-steering systems, **including standards.**
- Standards content must be
- **formalized/structured**
  - **explicit/error-free**
  - **context-based/context-sensitive**
  - **granularly addressable**
- to be executable by machine and software
- **SMART Standards**



**Figure 28:** AI usage features according to the “Utility Model”



The different AI usage features in the level model (see [Chapter 5.2.3](#)) are adapted in [Figure 28](#) as an example. In the project the scenarios according to Level 4 (see [Annex 11.4.3](#)) are to be aimed at. Level 2 and 3 already allow the use of granular information, see [Annex 11.4.1](#).

### 5.2.5 New design of standards for AI application processes

The procedures for providing granular standards information will vary:

- **Technological approach** (see details in [Annex 11.4.1](#)): Existing standards documents are automatically indexed in post-processing without subject and number limitations and are automatically provided in granular “addressable” information units using semantic methods. The indexing accuracy is currently about 80 % compared to intellectually granularly prepared documents and thus meets the requirements of qualified users who can evaluate the disassembled information offer professionally. But for downstream AI application processes this means: A validation of the accuracy of the partial information must be integrated. The drivers of this approach are “content management” and “content delivery”. Results in **Level 3** (with above mentioned limitations) **based on Level 2** are achievable.
- **Bottom up approach** (see details in [Annex 11.4.2](#)): When digitizing standards, a distinction can be made between a top-down and a bottom-up approach. Both approaches deal with questions of modularization, modelling and management of future standard content, but from different perspectives. Here, the top-down approach is characterized by the redesign of the actual standardization process and the question of how future digital standards must be structured, whereas the bottom-up

approach deals with the transfer of already existing standard contents (“restructuring”) into a machine-executable knowledge representation. The development of smart standards requires both a top-down and a bottom-up approach. The drivers of the bottom-up approach are “content management and delivery” and “content usage”. **Level 3 and Level 4 results** are achievable for defined, delimited areas of application.

- **Top down approach** (see details in [Annex 11.4.3](#)): There can only be one reference document or “reference content” of the standard and this is the content that has been checked and approved by the responsible standards body, the “primary content”. As a rule, laws or contracts refer only to these and only this primary content is relevant in serious cases. So that the machine-readable standard content can also be primary content, the acquisition of the human-generated and -readable linguistic standard content must be carried out in preprocessing (in the sense of the standards development process) on the basis of a structure that allows the language, including the semantics it contains, to be unambiguously transformed into a machine-readable data structure (e.g. ontology) and vice versa. The drivers of this approach are “content creation” and “content usage”. **Level 4 results** are achievable.

#### Processing sequence

The different approaches can and should be pursued in parallel. The technology approach provides faster insights than can be used in the other approaches. In addition, the first – economically viable – customer solutions or prototypes and demonstrators are quickly developed, so that practical experience can be fed back. The bottom-up approach cannot be suitable for structuring the very large and constantly growing worldwide stock of standards to the highest level of quality. But in this procedure the idea is also to start purposefully in

order to gain experience. However, the post-processing of standards can be economical for concrete fields of application. The “top class” for the pronounced goal of achieving SMART standards with the highest quality requirements for AI application processes can only be the pursuit and implementation of a top-down method (preprocessing). The effort for this will be very high.

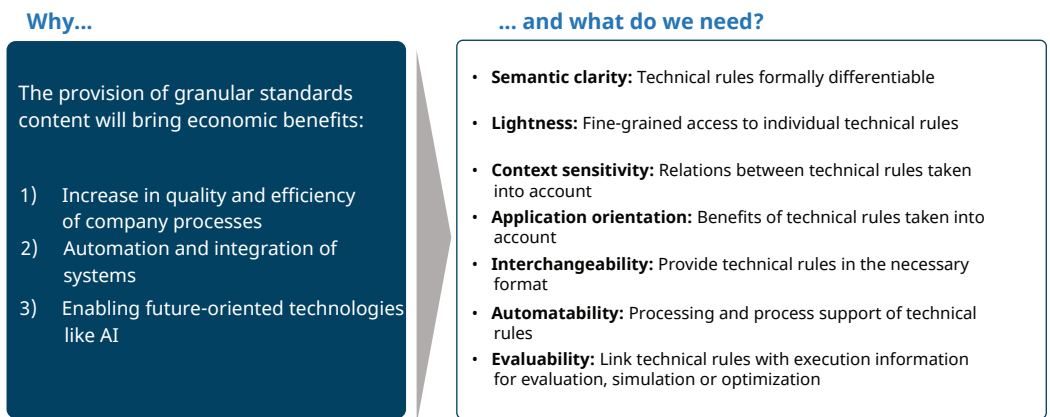
**Economic benefits**

The economic benefits of standardization are quantified in some countries. In Germany standardization saves the economy 17 billion euros per year [11]. In France, standardization contributes directly to the improvement of the gross domestic product, the effect of which is estimated to average over 5 billion euros per year. In the UK, 28,4% of the annual GDP growth can be attributed to standards, or 9 billion euros. The economic benefits of SMART standards have not yet been quantified and can only be described qualitatively, see Figure 29 [259].

Exactly such an aggregated statement (which admittedly is also bold) and the derivation of it are missing for the new approach SMART Standards. Typical questions are: What share of the benefits from SMART standards falls to the 17 billion euros mentioned today? Would we lose benefits if we do not deal with SMART standards? Or would benefits be added to the 17 billion euros?

Within the framework of the project, an economic evaluation must be made with regard to costs, benefits, implementation period, quality, etc. of the various approaches. Afterwards or while accompanying the project, a prioritization of the procedures can be made.

**Figure 29:** Prerequisites and benefits of SMART standards



**5.2.6 Summary and Outlook**

In this Chapter 5.2 “SMART Standards – New design of standards for AI application processes” a systematic approach to the development of SMART standards and information models for mapping standards and specifications is described. Both are not yet available today – but both are an important prerequisite for providing AI-based application models with reliable standard-compliant granular information.

With everything that the user can imagine today and in the future, ultimately, the aim is to develop downstream AI application processes, at the end of which systems are created that are capable of providing answers to user or application questions based on formalized and modelled expertise from standards and specifications and logical conclusions drawn from them.

In this paper numerous requirements with AI relevance are described, which can be derived from the SMART standards project. The implementation of these requirements raises the benefits of standardization to a significantly higher level.





**6**

**Overview of relevant documents, activities and committees on AI**

This chapter provides an overview of the essential standards and specifications (6.1 and 6.2), ongoing standardization activities (6.3) and standardization committees (6.4). In the

following tables the most important documents are listed and supplemented by additional information. The lists make no claim to completeness.

## 6.1 Published standards and specifications on AI

Table 10 lists existing standards and specifications that deal explicitly with AI applications. Neither the table as a whole nor the allocation to the main topics make any claim to completeness.

**Table 10:** Existing standards and specifications on AI

Document	Topic	Title	Brief description with possible relevance to AI	Relevance to topic						
				Principles (4.1)	Ethics/Responsible AI (4.2)	Quality, conformity assessment and certification (4.3)	IT Security in AI systems (4.4)	Industrial automation (4.5)	Mobility and logistics (4.6)	AI in medicine (4.7)
ISO/IEC 24028 [261]	AI principles	Information technology – Artificial intelligence – Overview of trustworthiness in artificial intelligence	Technical report on the trustworthiness of AI systems	x	x		x			x
ITU-T Y.3170 [154]	Service quality of future networks	Requirements for machine learning – based quality of service assurance for the IMT-2020 Network	Requirements for data collection, preparation and modelling with regard to quality of service and quality of experience (see also 4.3.1.3)			x				x
ITU-T Y.3173 [155]	Evaluation of the capabilities of future networks	Framework for evaluating intelligence level of future networks including IMT-2020	Estimating AI capabilities of networks (see also 4.3.1.3)			x				x
ETSI TS 103 296 [152]	Emotion detection	Speech and Multimedia Transmission Quality (STQ); Requirements for Emotion Detectors used for Telecommunication Measurement Applications; Detectors for written text and spoken speech	Requirements and characteristics relating to emotion detection (see also 4.3.1.3)		x	x				x

Document	Topic	Title	Brief description with possible relevance to AI	Relevance to topic							
				Principles (4.1)	Ethics/Responsible AI (4.2)	Quality, conformity assessment and certification (4.3)	IT Security in AI systems (4.4)	Industrial automation (4.5)	Mobility and logistics (4.6)	AI in medicine (4.7)	
ETSI TS 103 195-2 [153]	Network architecture	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	Autonomous systems: Requirements and use cases (see also 4.3.1.3)			x					x
ETSI GR ENI 004 [263]	AI terminology	Terminology for Main Concepts in ENI	Group Report (GR) on the terminology of networks with AI elements in the context of Experiential Networked Intelligence (ENI)	x							
ETSI GR NFV 003 [264]	AI terminology	Terminology for Main Concepts in NFV	Report (GR) on the terminology of networks with AI elements in the context of Network Functions Virtualisation (NFV)	x							
DIN SPEC 92001-1 [87]	AI life cycle	Artificial Intelligence – Life Cycle Processes and Quality Requirements – Part 1: Quality Meta Model	Concrete relation to AI: → Quality Meta Model Relation to ISO/IEC 12207 Life cycle model → Differentiation into risk classes: “low” and “high risk” AI modules → Quality columns: functionality and performance, robustness and comprehensibility → AI quality depends on: Design of the AI model and data quality → Risk management along the entire life cycle is recommended (see also 4.1.2.3, 4.3.1.3, 4.4.2.3 and 4.7.2)		x	x	x				x

Document	Topic	Title	Brief description with possible relevance to AI	Relevance to topic						
				Principles (4.1)	Ethics/Responsible AI (4.2)	Quality, conformity assessment and certification (4.3)	IT Security in AI systems (4.4)	Industrial automation (4.5)	Mobility and logistics (4.6)	AI in medicine (4.7)
DIN SPEC 92001-2 [318]	Life cycle of AI	Artificial Intelligence – Life Cycle Processes and Quality Requirements – Part 2: Robustness	AI-specific requirements with regard to robustness, especially regarding adversarial robustness and corruption robustness			x	x	x		
DIN SPEC 13266 [151]	Guideline for deep learning systems	Guideline for the development of deep learning image recognition systems	Procedure for data collection, structuring of data for learning AI image recognition, process structure of learning experiments and quality assurance (see also 4.3.1.3 and 4.7.3)	x		x	x	x		x
IEEE 7010-2020 [156]	Impact of autonomous systems on humans	Recommended Practice for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-being	Evaluation scheme of autonomous systems with regard to effects on human well-being (see also 4.3.1.3)	x	x	x				x
UL 4600 [157]	Evaluation of autonomous vehicles	Standard for the Evaluation of Autonomous Products	Covers safety principles, risk reduction, tools, techniques and life cycle processes for the development and evaluation of autonomous vehicles. Compatible with ISO/PAS 21448 and ISO 26262 (see also 4.3.1.3 and 4.6.1)		x	x			x	

## 6.2 Published standards and specifications with relevance to AI

Table 11 gives an overview of standards and specifications that do not yet make detailed statements about the application of AI components but are particularly relevant for AI standardization or AI application. Neither the table as a whole nor the allocation to the main topics make any claim to completeness.

**Table 11:** General standards and specifications with relevance to AI application

Document	Topic	Title	Brief description with possible relevance to AI	Relevance to topic						
				Principles (4.1)	Ethics/Responsible AI (4.2)	Quality, conformity assessment and certification (4.3)	IT Security in AI systems (4.4)	Industrial automation (4.5)	Mobility and logistics (4.6)	AI in medicine (4.7)
ISO/IEC/IEEE 12207 [58]	Life cycle of software	Systems and software engineering – Software life cycle processes	<p>→ Description of processes of the life cycle (idea generation to decommissioning) and their relationships to each other on an abstract level</p> <p>→ No specification of a life cycle model or development method (see also 4.1.2.3 and 4.3.1.2)</p>	x	x	x		x		
ISO/IEC/IEEE 29119 [265]–[269]	Software tests	Software Testing	<p>29119-1: Concepts &amp; Definitions</p> <p>29119-2: Test Processes</p> <p>29119-3: Test Documentation</p> <p>29119-4: Test Techniques</p> <p>29119-5: Keyword Driven Testing</p>			x		x		
ISO/IEC 15408 [48]–[50]	Security techniques	Information technology – Security techniques – Evaluation criteria for IT security	<p>Defines the Common Criteria (CC), 7 Evaluation Assurance Levels (EAL), 11 function classes, 7 organizational classes (see also 4.1.2.2, 4.1.2.3).</p> <p>Parts 1 to 3 have been published, Parts 4 to 5 are in development (see 6.3).</p>		x	x	x	x		x
ISO/IEC 17000ff [38]–[44]	Conformity assessment	Conformity assessment	<p>Family of standards on conformity assessment in general. Not AI-specific but forms the basis for AI conformity assessment (see also 4.1.2.1.5 and especially 4.3).</p>			x		x		x

Document	Topic	Title	Brief description with possible relevance to AI	Relevance to topic							
				Principles (4.1)	Ethics/Responsible AI (4.2)	Quality, conformity assessment and certification (4.3)	IT Security in AI systems (4.4)	Industrial automation (4.5)	Mobility and logistics (4.6)	AI in medicine (4.7)	
ISO/IEC 18045 [51]	Security techniques	Information technology – Security techniques – Methodology for IT security evaluation	Methodology for the evaluation of IT security on the basis of the CC (“Evaluation methodology”) (see 4.1.2.2 and 4.4.1.3)				x	x			
ISO/IEC 20546 [34]	Big Data	Information technology – Big data – Overview and vocabulary	Specifies terminology for big data (see also 4.1.1 and 4.3.1.2)	x		x					
ISO/IEC TR 20547-1 [270]	Big Data	Information technology – Big data reference architecture – Part 1: Framework and application process	Reference architecture for big data: processes	x		x					
ISO/IEC TR 20547-2 [149]	Big Data	Information technology – Big data reference architecture – Part 2: Use cases and derived requirements	Reference architecture for big data: use cases (see also 4.3.1.2)	x		x					x
ISO/IEC 20547-3 [271]	Big Data	Information technology – Big data reference architecture – Part 3: Reference architecture	Reference architecture for big data: terminology and concepts	x		x	x				
ISO/IEC TR 20547-5 [150]	Big Data	Information technology – Big data reference architecture – Part 5: Standards roadmap	Overview of standards relevant to big data (see also 4.3.1.2)			x		x			x
ISO/IEC 25000 [272]	Software quality	Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – Guide to SQuaRE	→ Guidelines for quality criteria and the evaluation of software products → Definition of the SQuaRE model			x	x	x			x

Document	Topic	Title	Brief description with possible relevance to AI	Relevance to topic							
				Principles (4.1)	Ethics/Responsible AI (4.2)	Quality, conformity assessment and certification (4.3)	IT Security in AI systems (4.4)	Industrial automation (4.5)	Mobility and logistics (4.6)	AI in medicine (4.7)	
ISO/IEC 25010 [146]	Software quality	Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – System and software quality models	<p>Defines quality criteria</p> <ul style="list-style-type: none"> <li>→ Functionality: Correctness, appropriateness, completeness</li> <li>→ Reliability: Maturity, fault tolerance, recoverability</li> <li>→ Usability: Understandability, operability, learnability, robustness</li> <li>→ Efficiency: economic efficiency, behaviour over time, consumption pattern</li> <li>→ Maintainability: Analysability, modifiability, stability, testability</li> <li>→ Portability: Adaptability, installability, conformity, replaceability</li> <li>→ Security: integrity, confidentiality, authenticity, accountability, non-repudiation</li> <li>→ Compatibility: interoperability (can be used to draw up the specification and test cases (see also 4.3.1))</li> </ul>			x	x	x			
ISO/IEC 25012 [89]	Data quality	Software engineering – Software product Quality Requirements and Evaluation (SQuaRE) – Data quality model	<p>Quality of the data product:</p> <ul style="list-style-type: none"> <li>→ inherent data quality (accuracy, completeness, consistency, credibility, currentness, accessibility, compliance, confidentiality, efficiency)</li> <li>→ System-dependent data quality ( availability, portability, recoverability, precision, traceability, understandability) (see also 4.1.2.3 and 4.3.1.2)</li> </ul>			x	x				

Document	Topic	Title	Brief description with possible relevance to AI	Relevance to topic						
				Principles (4.1)	Ethics/Responsible AI (4.2)	Quality, conformity assessment and certification (4.3)	IT Security in AI systems (4.4)	Industrial automation (4.5)	Mobility and logistics (4.6)	AI in medicine (4.7)
ISO/IEC 25020 [273]	Software quality	Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – Quality measurement framework	Guidelines for the selection, application and creation of quality characteristics			x	x			
ISO/IEC 25021 [274]	Software quality	Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – Quality measure elements	Quality criteria for software development			x	x			
ISO/IEC 25024 [275]	Data quality	Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – Measurement of data quality	Quality criteria and evaluation for secure software development			x	x			
ISO/IEC 27000ff [71]–[78], [122], [163], [210], [276]	Security techniques	Information technology – Security techniques	Family of standards on information security management systems (ISMS) with a set of substandards on various topics – e.g. guidelines for ISMS audit, data security; ISMS for health care, etc. and others → ISO/IEC 27034 Secure software development (see also 4.1.2.3 and 4.3.1.2) → ISO/IEC 27005 Risk Management (see also 4.4.2.3) → ISO/IEC 27701 Privacy information management (see also 4.3.2.3.2.3)	x	x	x				
ISO/IEC 29100ff [212], [277]	Security techniques	Information technology – Security techniques – Privacy framework	Standards family on data protection (Privacy framework), e.g. ISO/IEC 29134 privacy impact assessment (risk assessment)				x			
ISO/IEC 33063 [278]	Software tests	Information technology – Process assessment – Process assessment model for software testing	Guidelines for the definition and evaluation of criteria of process capability in manufacturing			x				



Document	Topic	Title	Brief description with possible relevance to AI	Relevance to topic						
				Principles (4.1)	Ethics/Responsible AI (4.2)	Quality, conformity assessment and certification (4.3)	IT Security in AI systems (4.4)	Industrial automation (4.5)	Mobility and logistics (4.6)	AI in medicine (4.7)
ISO 12100 [124], [125]	Safety of machinery	Safety of machinery – General principles for design – Risk assessment and risk reduction	Terminology and methodology as well as general guidelines for risk assessment and risk reduction in the manufacture of safe machinery  In <b>Chapter 6</b> : Statements on safety functions implemented by programmable electronic controllers (see also 4.2.2.4 and 4.3.1.2)		x	x	x			
ISO 13849 [126], [127]	Safety of machinery	Safety of machinery – Safety-related parts of control systems	Principles of design and integration of safety-related parts of control systems and programmable electronic systems (see also 4.2.2.4)		x	x	x			
ISO 14971 [128]	Risk management	Medical devices – Application of risk management to medical devices	Terminology, principles and process for risk management of medical devices, including software as a medical device  Example of a standard according to which safety-relevant AI systems are currently designed, see also Ethics (4.2.2.4)		x	x				x
ISO/PAS 21448 [148]	Safety of intended functionality	Road vehicles – Safety of the intended functionality	Safety of the intended functionality (SOTIF)  → Considers inappropriate risk due to hazards caused by functional deficiencies of the intended functionality or by reasonably foreseeable misuse by persons  → Performance restrictions can also be assigned to the environment and communication (see also 4.3.1.2)			x	x	x	x	

Document	Topic	Title	Brief description with possible relevance to AI	Relevance to topic						
				Principles (4.1)	Ethics/Responsible AI (4.2)	Quality, conformity assessment and certification (4.3)	IT Security in AI systems (4.4)	Industrial automation (4.5)	Mobility and logistics (4.6)	AI in medicine (4.7)
ISO/TR 22100 [279]–[281]	Document overview: Risk reduction for machinery	Safety of machinery – Relationship with ISO 12100	Targeted selection of the various types of ISO standards on machine safety 4 parts published, further parts in development (see 6.3)			x	x	x		
ISO 23412 [282]	Logistics	Indirect, temperature-controlled refrigerated delivery services – Land transport of parcels with intermediate transfer	Concentrates on the technical and organizational implementation of the transport of refrigerated goods, but can be seen as a cornerstone for automatic distribution of goods and is therefore relevant for AI applications (see also 4.6.1)						x	
ISO 25119 [283]–[286]	Functional safety	Tractors and machinery for agriculture and forestry – Safety-related parts of control systems	Safety by Design, development, conception and production			x	x		x	x
ISO 26262 [59]–[70]	Functional safety	Road vehicles – Functional safety	Management of functional safety Concept phase Product development: System level Product development: Hardware level Product development: Software level Production, operation and decommissioning Supporting processes ASIL and safety-oriented analyses (see also 4.1.2.3)			x	x		x	
ISO 31000 [93]	Risk assessment	Risk management – Guidelines	General, not AI-specific guidelines to risk management approach for handling any type of risk, not industry- or sector-specific. Basis for ISO/IEC 23894 on risk management for AI. (see also 4.1.3, 4.2.2.4, 4.4.2.3)		x	x	x			x

Document	Topic	Title	Brief description with possible relevance to AI	Relevance to topic							
				Principles (4.1)	Ethics/Responsible AI (4.2)	Quality, conformity assessment and certification (4.3)	IT Security in AI systems (4.4)	Industrial automation (4.5)	Mobility and logistics (4.6)	AI in medicine (4.7)	
IEC 60601-1-4 [287]	Medical devices	Medical electrical equipment – Part 1-4: General requirements for safety – Collateral Standard: Programmable electrical medical systems	Requirements on safety, testing and guidelines for programmable electrical medical systems.		x	x					x
IEC 61508 [86]–[70]	Functional safety of systems	Functional safety of electrical/electronic/programmable electronic safety-related systems	IEC 61508-3: Requirements on software: Artificial intelligence for error correction explicitly not recommended for SIL 2 and higher IEC 61508-5: Examples for determining the safety integrity level (see also 4.1.2.3, 4.3.1.2, 4.4.2.3 and 4.5.2.3)	x		x	x				x
IEC 61511 [211], [288], [289]	Functional safety of process control technology	Functional safety – Safety instrumented systems for the process industry sector	Part 1: General, terms and definitions, requirements on systems, software and hardware (see also 4.4.2.3)			x	x				
IEC 61513 [290]	Requirements on control systems and devices	Nuclear power plants – Instrumentation and control important to safety – General requirements for systems	Safety life cycle concept for the entire control architecture and individual systems		x	x	x				
IEC 62061 [129]	Functional safety of control systems	Safety of machinery – Functional safety of safety-related electrical, electronic and programmable electronic control systems	Selection and design of a safety-related electrical, electronic and programmable electronic control system (SRECS) and approach to risk assessment and determination of the safety integrity level (SIL) (see also 4.2.2.4)		x	x	x				
IEC 62304 [291]	Software life cycle	Medical device software – Software life cycle processes	Development and maintenance of medical device software		x	x	x				x

Document	Topic	Title	Brief description with possible relevance to AI	Relevance to topic						
				Principles (4.1)	Ethics/Responsible AI (4.2)	Quality, conformity assessment and certification (4.3)	IT Security in AI systems (4.4)	Industrial automation (4.5)	Mobility and logistics (4.6)	AI in medicine (4.7)
IEC 62443 [199]–[209]	IT security	Industrial communication networks – Network and system security	Series of standards in 13 parts defines terminology (IEC TS 62443-1-1) and requirements, e.g. for the IT security program of service providers (IEC 62443-2-4), the life cycle for secure product development (IEC 62443-4-1) and security levels (IEC 62443-3-3) (see also 4.4.1.3)				x			
DIN EN 50128 [292]	Safety-relevant software for railway applications	Railway applications – Communication, signalling and processing systems – Software for railway control and protection systems	Methods, principles and measures for software safety			x	x		x	
ETSI TR 101 583 [175]	Security tests	Methods for Testing and Specification (MTS); Security Testing; Basic Terminology	Enumeration and explanation of relevant methods and approaches for security testing, such as risk analysis and risk-based security testing, functional testing of security functions, performance testing, robustness testing and penetration testing (see also 4.3.2.3.2.4)		x	x	x	x		
IEEE 1012-2016 [293]	Validation of hardware and software	Standard for System, Software, and Hardware Verification and Validation	Verification during product development as to whether requirements are met			x		x		

### 6.3 Current standardization activities on AI

Table 12 gives information on current standardization projects. Neither the table as a whole nor the allocation to the main topics make any claim to completeness.

**Table 12:** Overview of current AI-relevant standardization projects

Document	Topic	Title	Brief description with possible relevance to AI	Relevance to topic							
				Principles (4.1)	Ethics/Responsible AI (4.2)	Quality, conformity assessment and certification (4.3)	IT Security in AI systems (4.4)	Industrial automation (4.5)	Mobility and logistics (4.6)	AI in medicine (4.7)	
ISO/IEC WD TS 4213	Assessment of AI systems	Information technology – Artificial Intelligence – Assessment of machine learning classification performance	Metrics for the performance capability of AI	x		x		x			
ISO/IEC NP 5059	Software quality	Software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) Quality Model for AI-based systems	Quality assessment for AI-based systems (see also 4.1.1 and 4.3.1.4)	x		x		x			
ISO/IEC WD 5259-1	Data quality	Data quality for analytics and ML – Part 1: Overview, terminology, and examples	Data quality management for machine learning: Overview, terminology and examples	x		x		x			
ISO/IEC WD 5259-3	Data quality	Data quality for analytics and ML – Part 3: Data Quality Management Requirements and Guidelines	Data quality management for machine learning: Requirements and guidelines	x		x		x			
ISO/IEC WD 5259-4	Data quality	Data quality for analytics and ML – Part 4: Data quality process framework	Data quality management for machine learning: Processes	x		x		x			
ISO/IEC NP 5338	Development of AI systems	Information technology – Artificial intelligence – AI system life cycle processes	Terminology standard on life cycle process of AI systems (in voting phase)	x		x	x				
ISO/IEC NP 5339	Application guidelines	Information Technology – Artificial Intelligence – Guidelines for AI Applications	Guidelines for application of AI systems (in voting phase)	x	x	x					
ISO/IEC NP 5392	AI systems	Information technology – Artificial intelligence – Reference Architecture of Knowledge Engineering	Reference architecture for knowledge-based systems	x		x	x				

Document	Topic	Title	Brief description with possible relevance to AI	Relevance to topic							
				Principles (4.1)	Ethics/Responsible AI (4.2)	Quality, conformity assessment and certification (4.3)	IT Security in AI systems (4.4)	Industrial automation (4.5)	Mobility and logistics (4.6)	AI in medicine (4.7)	
ISO/IEC NP 5394	AI principles	Information Technology – Artificial intelligence – Management System	Management system standard for AI (see 4.1.2.2)	x		x	x				
ISO/IEC AWI TR 5469	Functional safety and AI	Functional Safety and AI Systems	The document will describe properties, relevant risk factors, usable methods and processes for the application of AI in safety-relevant functions, for the application of safety-relevant functions for the control of AI systems and for the application of AI in the development of safety-relevant functions. It is being developed in coordination with IEC/SC 65 A (the standardization group responsible for IEC 61508).			x	x	x			
ISO/IEC 15408	Security techniques	Information technology – Security techniques – Evaluation criteria for IT security	Defines the Common Criteria (CC), 7 Evaluation Assurance Levels (EAL), 11 function classes, 7 organizational classes (see also 4.1.2.2, 4.4.2.3). Parts 1 to 3 have been published (see 6.2), Parts 4 and 5 are in development				x	x			
ISO/IEC FDIS 20547-4	Big Data	Information technology – Big data reference architecture – Part 4: Security and privacy	Reference architecture for big data			x	x	x			x
ISO/IEC CD 22989	AI terminology	Artificial intelligence – Concepts and terminology	Basic standard describing concepts and terminology for artificial intelligence (see also 4.1.1 and 4.6.2.1)	x		x		x			
ISO/IEC CD 23053	Machine learning	Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)	Describes a terminological framework for machine learning (see also 4.1.1)	x				x			

Document	Topic	Title	Brief description with possible relevance to AI	Relevance to topic							
				Principles (4.1)	Ethics/Responsible AI (4.2)	Quality, conformity assessment and certification (4.3)	IT Security in AI systems (4.4)	Industrial automation (4.5)	Mobility and logistics (4.6)	AI in medicine (4.7)	
ISO/IEC CD 23894	AI risk management	Information Technology – Artificial Intelligence – Risk Management	Contains risk management guidelines for the development and use of AI systems (see also 4.1.1, 4.1.3, 4.2.3, 4.4.2.3)	x		x	x	x			x
ISO/IEC AWI TR 24027	AI principles	Information technology – Artificial Intelligence (AI) – Bias in AI systems and AI aided decision making	Technical report describing “bias” in AI systems (see also 4.1.1)	x	x			x			
ISO/IEC NP 24029	AI robustness	Artificial Intelligence (AI) – Assessment of the robustness of neural networks	ISO/IEC CD TR 24029-1: Overview ISO/IEC AWI 24029-2: Formal methods methodology (see also 4.5.2.3)			x		x			x
ISO/IEC CD TR 24030	Applications	Information technology – Artificial Intelligence (AI) – Use cases	Collection of use cases for AI systems	x			x	x			
ISO/IEC AWI TR 24368	Ethics	Information technology – Artificial intelligence – Overview of ethical and societal concerns	Technical report on ethical and societal concerns relating to AI (see also 4.1.1)	x		x					
ISO/IEC AWI TR 24372	AI principles	Information technology – Artificial intelligence (AI) – Overview of computational approaches for AI systems	Technical report on AI methods	x		x		x			
ISO/IEC WD TS 24462	Trustworthiness	Ontology for ICT Trustworthiness Assessment	New project for a Technical Specification. Being developed in ISO/IEC JTC 1/WG 13 “Trustworthiness”			x	x	x			
ISO/IEC AWI 24668	Big Data	Information technology – Artificial intelligence – Process management framework for Big data analytics	Management of data analyses for big data	x		x					
ISO/IEC CD 38507	Governance	Information technology – Governance of IT – Governance implications of the use of artificial intelligence by organizations	Deals with organizational governance in connection with AI (see also 4.1.1)	x		x					

Document	Topic	Title	Brief description with possible relevance to AI	Relevance to topic						
				Principles (4.1)	Ethics/Responsible AI (4.2)	Quality, conformity assessment and certification (4.3)	IT Security in AI systems (4.4)	Industrial automation (4.5)	Mobility and logistics (4.6)	AI in medicine (4.7)
ISO/SAE 21434	IT security	Road vehicles – Cybersecurity engineering	Defines terminology and the most important aspects of IT security				x			
ISO/CD TR 4804	IT security/safety	Road vehicles – Safety and cybersecurity for automated driving systems – Design, verification and validation methods	Work on AI in road vehicles. Is being developed in ISO/TC 22 “Road vehicles” (see also 4.6.1)				x	x	x	
ISO/CD TR 22100-5	Machine safety and AI	Safety of machinery – Relationship with ISO 12100 – Part 5: Implications of embedded Artificial Intelligence-machine learning	The technical report describes how hazards arising from the use of ML systems in machines should be considered in the risk assessment process.			x	x	x		
ISO/AWI 24089	IT security	Road vehicles – Software update engineering	New specification in development			x	x	x		
ITU-T F.AI-DLFE	Evaluation of software on the basis of deep learning	Deep Learning Software Framework Evaluation Methodology	Requirements on the architectures of deep learning			x		x		
ITU-T F.AI-DLPB	Metrics and evaluation of neural networks	Metrics and evaluation methods for deep neural network processor benchmark	Evaluation scheme for deep learning with regard to inference, training, application, network and processor			x		x		
ITU-T F.VS-AIMC	Requirements for data transmission	Use cases and requirements for multimedia communication enabled vehicle systems using artificial intelligence	Network requirements with regard to prediction, planning, human-machine interaction and training of models			x		x		x
ITU-T Y.qos-ml-arc	Service quality of future networks	Architecture of machine learning based QoS assurance for the IMT-2020 network	Service quality in future networks with regard to machine learning			x		x		
ETSI DGS SAI 003	AI security tests	Securing Artificial Intelligence (SAI); Security Testing of AI	Guidelines for security tests for AI components The focus is on data for machine learning			x	x	x		



Document	Topic	Title	Brief description with possible relevance to AI	Relevance to topic							
				Principles (4.1)	Ethics/Responsible AI (4.2)	Quality, conformity assessment and certification (4.3)	IT Security in AI systems (4.4)	Industrial automation (4.5)	Mobility and logistics (4.6)	AI in medicine (4.7)	
ETSI DGR SAI 002	AI training data quality	Securing Artificial Intelligence (SAI); Data Supply Chain Report	Overview of existing procedures for data collection, rules for data handling, Identification of standardization needs			x	x	x			
ETSI DTR INT 008 (TR 103 821)	AI tests	Artificial Intelligence (AI) in Test Systems and Testing AI models: Testing of AI, with Definitions of Quality Metrics	Test framework for systems of network automation such as ETSI GANA (Generic Autonomous Networking Architecture)			x		x			
IEEE P2801	Data quality	Recommended Practice for the Quality Management of Datasets for Medical Artificial Intelligence	QM system for data preparation for AI medical devices			x		x			x
IEEE P2802	Terminology: Evaluation of AI safety aspects	Standard for the Performance and Safety Evaluation of Artificial Intelligence Based Medical Device: Terminology	Safety, risk, effectiveness and QM			x	x	x			x
IEEE P2846	Mobility	A Formal Model for Safety Considerations in Automated Vehicle Decision Making	Technology neutral mathematical model and test method for automatic decision-making in vehicles (see 4.6.1)			x				x	
IEEE P3333.1.3	Assessment on the basis of deep learning	Standard for the Deep Learning-Based Assessment of Visual Experience Based on Human Factors	Assessment of subjective and objective user-friendliness via deep learning	x		x		x			x
DIN SPEC 2343	Data transmission	Transmission of language-based data between artificial intelligences – Specification of parameters and formats	Format for transferring speech data between different ecosystems for industrial users, open source communities and private users with a focus on interoperability and traceability			x		x			
DIN SPEC 91426	Video analysis	Quality requirements for video-based methods of personnel selection	Procedure to avoid mistakes, prevent discrimination and increase the prognostic validity of digital recruitment procedures			x		x			

Document	Topic	Title	Brief description with possible relevance to AI	Relevance to topic						
				Principles (4.1)	Ethics/Responsible AI (4.2)	Quality, conformity assessment and certification (4.3)	IT Security in AI systems (4.4)	Industrial automation (4.5)	Mobility and logistics (4.6)	AI in medicine (4.7)
NISTIR 8269	IT security	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		
VDE AR 2842-61	Trustworthiness	Development and trustworthiness of autonomous/cognitive systems	Defines a general framework for the development of trusted solutions and trusted autonomous/cognitive systems, including requirements for the subsequent phases of the product life cycle (e.g. production, marketing & sales, operation & maintenance, decommissioning & repair). Defines a reference life cycle by analogy with the major functional safety standards (i.e. IEC 61508) as a unified approach to achieve and maintain the overall performance of the solution and the intended behaviour and reliability of the autonomous/cognitive system. Furthermore, this could lead to a basis for the qualification and conformity assessment of solutions based on autonomous/cognitive systems including elements of artificial intelligence. (see also 4.5.1)		x	x	x	x	x	x

## 6.4 Committees on AI

The following table gives an overview of the most important AI standardization committees. Neither the table as a whole nor the allocation to the topics of the committees (see the introductory paragraph to [Chapter 4](#)) claim to be complete.

**Table 13:** Overview of the most important AI standardization committees

	Committee	Mirror committee <sup>a</sup>	Relevance to topic						
			Principles (4.1)	Ethics/Responsible AI (4.2)	Quality, conformity assessment and certification (4.3)	IT Security in AI systems (4.4)	Industrial automation (4.5)	Mobility and logistics (4.6)	AI in medicine (4.7)
International	ISO/IEC JTC 1/SC 7 “Software and system engineering”	NA 043-01-07 AA	x		x		x		
	ISO/IEC JTC 1/SC 27 “Information security, cybersecurity and privacy protection”	NA 043-01-27 AA	x	x	x	x	x		
	ISO/IEC JTC 1/SC 29 “Coding of audio, picture, multimedia and hypermedia information”	NA 043-01-29 AA					x		x
	ISO/IEC JTC 1/SC 31 “Automatic identification and data capture”	NA 043-01-31 AA					x		
	ISO/IEC JTC 1/SC 37 “Biometrics”	NA 043-01-37 AA					x		x
	ISO/IEC JTC 1/SC 38 “Cloud management and distributed platforms”	NA 043-01-37 AA					x		
	ISO/IEC JTC 1/SC 40 “IT Service Management and IT Governance”	NA 043-01-40 AA	x		x		x		
	ISO/IEC JTC 1/SC 41 “Internet of things and related technologies”	NA 043-01-41 AA			x		x		
	ISO/IEC JTC 1/SC 42 “Artificial Intelligence”	NA 043-01-42 AA	x	x	x	x	x	x	
	ISO/TC 199 “Safety of machinery”	NA 095 BR	x	x	x		x		x
	ISO/TC 204 “Intelligent transport systems”/AG 1 “Big data and artificial intelligence”	NA 052-00-71 GA	x		x		x		
	ISO/TC 299 “Robotics”	NA 060-38-01 AA	x		x		x		x
	IEC SEG 10 “Ethics in Autonomous and Artificial Intelligence Applications”	DKE/TBINK AG			x		x		

Committee	Mirror committee <sup>a</sup>	Relevance to topic					
		Principles (4.1)	Ethics/Responsible AI (4.2)	Quality, conformity assessment and certification (4.3)	IT Security in AI systems (4.4)	Industrial automation (4.5)	Mobility and logistics (4.6)
IEC/TC 62 “Electrical equipment in medical practice”/AG SNAIG “Software Network and Artificial Intelligence advisory Group”				x		x	
IEC/TC 65 “Industrial process measurement – control and automation”/SC 65A “System aspects” (Liaison with ISO/IEC JTC 1/SC 42)				x	x	x	
IEC/TC 65 JWG23 “Usage of new Technologies”						x	
IEC/TC 65/WG 23 “Smart Manufacturing Framework and System Architecture”						x	
ITU-T FG – AI4AD “Focus Group on AI for autonomous and assisted driving”				x		x	x
ITU-T FG – AI4EE “Focus Group on Environmental Efficiency for Artificial Intelligence and other Emerging Technologies”		x				x	
ITU-T FG – AI4H “Focus Group on Artificial Intelligence for Health”		x		x		x	x
ITU-T FG – ML5G “Focus Group on Machine Learning for Future Networks including 5G”		x				x	
ITU-T SG 2 “Operational aspects”				x		x	x
ITU-T SG 5 “Environment and circular economy”		x		x		x	
ITU-T SG 13 “Future networks (& cloud)”		x				x	
ITU-T SG 16 “Multimedia”		x				x	
ITU-T SG 20 “IoT, smart cities & communities”		x				x	
European							
CEN-CENELEC Focus Group on Artificial Intelligence	NA 043-01-42 AA	x		x		x	
CEN/CLC/JTC 13 “Cybersecurity and Data Protection”	NA 043 BR-07 SO					x	x
ETSI ISG ENI “Experiential Network Intelligence”				x		x	x
ETSI ISG NFV “Network Function Virtualisation”				x		x	
ETSI ISG SAI “Securing Artificial Intelligence”		x		x	x	x	x
ETSI ISG ZSM “Zero touch network & Service Management”		x				x	
ETSI TC STQ “Speech and Multimedia Transmission Quality”		x		x		x	
ETSI TC Cyber “Cybersecurity”					x	x	

Committee	Mirror committee <sup>a</sup>	Relevance to topic						
		Principles (4.1)	Ethics/Responsible AI (4.2)	Quality, conformity assessment and certification (4.3)	IT Security in AI systems (4.4)	Industrial automation (4.5)	Mobility and logistics (4.6)	AI in medicine (4.7)
National	DIN NA 043-01 FB “Special Division Basic Standards of Information Technology” (with mirror committees to ISO/IEC JTC 1 as listed above), including NA 043-01-42 AA “Artificial Intelligence”	x		x		x		
	DKE/AK 801.0.8 “Specification and design of autonomous/cognitive systems”			x		x		
	DKE/AK 914.0.11 “Functional safety and artificial intelligence”			x	x	x		
	DKE/TBINK AG “Ethics and artificial intelligence”			x		x		x
	DIN SPEC 2343 “Transmission of language-based data between artificial intelligences – Specification of parameters and formats”			x		x		
	DIN SPEC 13266 “Guideline for the development of deep learning image recognition systems”			x		x		x
	DIN SPEC 92001 “Artificial Intelligence – Life Cycle Processes and Quality Requirements”	x		x	x	x		
	DIN SPEC 91426 “Quality requirements for video-based methods of personnel selection”			x		x		
Consortia	IETF NMRG “Network Management Research Group”	x						
	IETF COIN “Computing in the Network Proposed Research Group”	x						
	IETF ICNRG “Information-Centric Networking Research Group”	x						
	IETF TSVWG “Transport Area Working Group”	x						
	IEEE AIMDWG “Artificial Intelligence Medical Device Working Group”	x		x				x
	IEEE P7006 “Standard for Personal Data Artificial Intelligence (AI) Agent”			x				
	IEEE P7010 “Well-being Metrics for Autonomous and Intelligent Systems”	x		x				
	IEEE P7014 “Emulated Empathy in Autonomous and Intelligent Systems”							
	IEEE P3652.1 “Guide for Architectural Framework and Application of Federated Machine Learning”			x				
	IEEE P2841 “Framework and Process for Deep Learning Evaluation”			x				
	IEEE P3333.1.3 “Deep Learning-Based Assessment of Visual Experience Based on Human Factors”			x				x
IEEE P2807 “Framework of Knowledge Graphs”			x					

Committee	Mirror committee <sup>a</sup>	Relevance to topic					
		Principles (4.1)	Ethics/Responsible AI (4.2)	Quality, conformity assessment and certification (4.3)	IT Security in AI systems (4.4)	Industrial automation (4.5)	Mobility and logistics (4.6)
CSA “Working Group Artificial Intelligence”				x			
OGC “Artificial Intelligence in Geoinformatics Domain Working Group”							
OMG “Artificial Intelligence Platform level Task Force”				x			
W3C AI KR “Artificial Intelligence Knowledge Representation”				x			

a Titles of the DIN Standards Committees named:

NA 043 “DIN Standards Committee on Information Technology and selected IT Applications (NIA)”

NA 052 “DIN Standards Committee Road Vehicle Engineering (NAAutomobil)”

NA 060 “DIN Standards Committee Mechanical Engineering (NAM)”

NA 095 “DIN Standards Committee Safety Design Principles (NASG)”

NA 147 “DIN Standards Committee Quality Management, Statistics and Certification (NQSZ)”

NA 175 “DIN Standards Committee for Organizational Processes (NAOrg)”

7

## List of abbreviations

Abbreviation	Meaning
5G	Fifth generation mobile standard
AA	Arbeitsausschuss (Working Committee)
ACLU	American Civil Liberties Union
ACM	Association for Computing Machinery
ADM	Automated decision making/Algorithm decision making
AG	Arbeitsgruppe (Working Group)
AI	Artificial intelligence
AK	Arbeitskreis (Working Group)
ALKS	Automated Lane Keeping System
AML	Adversarial Machine Learning
AR (VDE-)	VDE-Anwendungsregel (application rule)
ASIL	Automotive Safety Integrity Level
Bitkom	Federal Association Information Technology, Telecommunications and New Media
BLEU	Bi-Lingual Evaluation Understudy
BMAS	Federal Ministry of Labour and Social Affairs
BMBF	Federal Ministry of Education and Research
BMI	Federal Ministry of the Interior, Building and Community
BMVI	Federal Ministry of Transport and Digital Infrastructure
BMWi	Federal Ministry for Economic Affairs and Technology
BR	Advisory Board
BSI	Federal Office for Information Security
BSI-KritisV	Regulation for the determination of critical infrastructures according to the BSI law
BVDW	German Association for the Digital Economy
CC	Common Criteria

Abbreviation	Meaning
CD	Committee Draft
CE	CE marking (Conformité Européenne)
CEN	Comité Européen de Normalisation, European Committee for Standardization
CENELEC	Comité Européen de Normalisation Électrotechnique, European Committee for Electrotechnical Standardization
CLC	Abbreviation for CENELEC used in committee designations
COM	EU Commission Communication
CSA	Cloud Security Alliance
DARPA	Defense Advanced Research Project Agency
DFKI	German Research Centre for Artificial Intelligence
DGR	Draft Group Report
DGS	Draft Group Specification
DIHK	German Chamber of Commerce and Industry
DIN	German Institute for Standardization
DIN SPEC	DIN Specification (consortial standard)
DKE	German Commission for Electrical, Electronic & Information Technologies
GDPR	EU General Data Protection Regulation, Regulation (EU) 2016/679
DSK	Data Protection Conference
EAL	Evaluation Assurance Level
E/E/PE	Electrical, electronic, programmable electronic
ENI	Experiential Networked Intelligence
ENISA	European Network and Information Security Agency
ETSI	European Telecommunications Standards Institute



Abbreviation	Meaning	Abbreviation	Meaning
EU	European Union	ISO	International Organization for Standardization
FG	Focus Group	IT	Information Technology
FMECA	Failure Mode and Effects and Critical Analysis	IT-SiG	German IT Security Act
GA	Joint Working Committee	ITU	International Telecommunication Union
GAIA-X	Project to build an efficient and competitive, secure and trustworthy data infrastructure for Europe	ITU-T	ITU Telecommunication Standardization Sector
GMA	VDI/VDE Society for Measurement and Automatic Control	JTC	Joint Technical Committee
GR	Group Report	JWG	Joint Working Group
AI HLEG	High Level Expert Group for AI	KBS	Conformity Assessment Body (CAB)
HR	Human Resources	KI	Artificial Intelligence (AI)
IAIS	Institute for Intelligent Analysis and Information Systems (Fraunhofer Institute)	SME	Small- and mid-sized enterprises
IBM	International Business Machines Corporation	KRITIS	Critical infrastructure within the meaning of the BSI-KritisV
ICT	Information and communication technology	LIME	Local Interpretable Model-Agnostic Explanations
IDC	International Data Corporation	MIP	Mixed Integer Linear Programming
IACS	Industrial Automation and Control Systems	ML	Machine Learning
IEC	International Electrotechnical Commission	MLP	Multi-Layer Perceptron
IEEE	Institute of Electrical and Electronics Engineers	MPG	Medical Devices Act
IETF	Internet Engineering Task Force	MSS	Management System Standard
IIRA	Industrial Internet Reference Architecture	MT	Machine translation
IKT	Informations- und Kommunikationstechnik	NA	Standards Committee (in DIN)
IMT-2020	International Mobile Telecommunications-2020	NASA	National Aeronautics and Space Administration
IoT	Internet of Things	NFV	Network Function Virtualisation
ISMS	Information Security Management Systems	NIST	National Institute of Standards and Technology, U.S. Department of Commerce
		NQDM	Draft standard for quality Data and Metadata (Fraunhofer Guidelines)

Abbreviation	Meaning
NRM	Standardization Roadmap
OECD	Organisation for Economic Co-operation and Development
OGC	Open Geospatial Consortium
OMG	Object Management Group
OPC	Open Platform Communications
OT	Operational IT
OWL	Ontology Web Language
PDW	Principle of Double Effect
PI4.0	Platform Industrie 4.0
PLS	Plattform Lernende Systeme (Platform Learning Systems)
PLT	Process Control Engineering
QM	Quality management
QoS	Quality of Service
SAI	Securing Artificial Intelligence
SafeTRANS	Safety in Transportation Systems
SAT	Satisfiability Theories
SC	Sub Committee
SCI4.0	Standardization Council Industrie 4.0
SEDRIS	Synthetic Environment Data Representation and Interchange Specification
SEG	Standardization Evaluation Group
SG	Study Group
SIL	Safety Integrity Level
SMS	Short Message Service
SMT	Satisfiability Modulo Theories
SO	Special committee
SPEC	DIN Specification (consortial standard)

Abbreviation	Meaning
SQuaRE	Systems and software Quality Requirements and Evaluation
TBINK	Technical Advisory Board International Coordination
TC	Technical Committee
TKG	Telecommunications Act
TMG	Telemedia Act
TR	Technical Report
TS	Technical Specification
TÜV	Technical inspection association
UL	UL LLC (Underwriters Laboratories)
UNECE	United Nation Economic Commission for Europe
VDE	Association for Electrical, Electronic & Information Technologies
VDI	Association of German Engineers
VDMA	German Engineering Federation
W3C	World Wide Web Consortium
WG	Working Group
WKIO	Werte, Kriterien, Indikatoren, Observablen (Values, criteria, indicators, observables)
WMA	World Medical Association
ZDH	German Confederation of Skilled Crafts
ZVEI	Central Association of Electrical Engineering and Electronics

8

## Sources and bibliography

- 
- [1] T. Heilmann, N. Schön, **NEUSTAAT: Politik und Staat müssen sich ändern. 64 Abgeordnete & Experten fangen bei sich selbst an – mit 103 Vorschlägen**. München: FinanzBuch, 2020.
- 
- [2] PwC, **Auswirkungen der Nutzung von künstlicher Intelligenz in Deutschland**. 2018 [Online]. Available at: <https://www.pwc.de/de/business-analytics/sizing-the-price-final-juni-2018.pdf> [Last retrieved: 17.08.2020].
- 
- [3] PwC, **Global Top 100 companies by market capitalization**. 2019 [Online]. Available at: <https://www.pwc.com/gx/en/audit-services/publications/assets/global-top-100-companies-2019.pdf> [Last retrieved: 17.08.2020].
- 
- [4] acatech (Ed.), **Künstliche Intelligenz in der Industrie, (acatech HORIZONTE)**. München, 2020.
- 
- [5] AI HLEG, **Ethics Guidelines for Trustworthy AI**. Brussels: European Commission, 2018 [Online]. Available at: <https://ec.europa.eu/futurium/en/ai-alliance-consultation> [Last retrieved: 13.08.2020].
- 
- [6] S. Palacio et al., **What do Deep Networks Like to See**. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018.
- 
- [7] Y. Gil, B. Selman, **A 20-Year Community Roadmap for Artificial Intelligence Research in the US. Computing Community Consortium (CCC) and Association for the Advancement of Artificial Intelligence (AAAI)**. 06.08.2019. [Online]. Available at: <https://cra.org/ccc/wp-content/uploads/sites/2/2019/08/Community-Roadmap-for-AI-Research.pdf> [last retrieved: 07.08.2020].
- 
- [8] W. Wahlster, **Künstliche Intelligenz versus menschliche Intelligenz**. Vorlesungsreihe 2017: Künstliche Intelligenz für den Menschen: Digitalisierung mit Verstand. Johannes Gutenberg Stiftungsprofessur. [Online]. Available at: [http://www.dfki.de/wwdata/Gutenberg\\_Stiftungsprofessur\\_Mainz\\_2017/Lernende\\_Maschinen.pdf](http://www.dfki.de/wwdata/Gutenberg_Stiftungsprofessur_Mainz_2017/Lernende_Maschinen.pdf) [Last retrieved: 11.08.2020].
- 
- [9] UBA, **Künstliche Intelligenz im Umweltbereich – Anwendungsbeispiele und Zukunftsperspektiven im Sinne der Nachhaltigkeit**. Dessau-Roßlau, 2019. [Online]. Available at: <https://www.umweltbundesamt.de/publikationen/kuenstliche-intelligenz-im-umweltbereich> [Last retrieved: 28.07.2020].
- 
- [10] Datenethikkommission der Bundesregierung, **Gutachten der Datenethikkommission**. Berlin: BMI, 2019. [Online]. Available at: [https://www.bmi.bund.de/SharedDocs/downloads/DE/publikationen/themen/it-digitalpolitik/gutachten-datenethikkommission.pdf?\\_\\_blob=publicationFile&v=6](https://www.bmi.bund.de/SharedDocs/downloads/DE/publikationen/themen/it-digitalpolitik/gutachten-datenethikkommission.pdf?__blob=publicationFile&v=6) [Last retrieved: 10.08.2020].
- 
- [11] K. Blind, A. Jungmittag, A. Mangelsdorf, **Der gesamtwirtschaftliche Nutzen der Normung – Eine Aktualisierung der DIN-Studie aus dem Jahr 2000**. Berlin: DIN, 2011. [Online]. Available at: <https://www.din.de/de/ueber-normen-und-standards/nutzen-fuer-die-wirtschaft> [Last retrieved: 07.08.2020].
- 
- [12] BMWi, **Strategie Künstliche Intelligenz der Bundesregierung**. Berlin, 2018. [Online]. Available at: [www.ki-strategie-deutschland.de](http://www.ki-strategie-deutschland.de) [Last retrieved: 13.08.2020].
- 
- [13] BMF, **Mit Zuversicht und voller Kraft aus der Krise, 3.6.2020**. [Online]. Available at: <https://www.bundesfinanzministerium.de/Content/DE/Standardartikel/Themen/Schlaglichter/Konjunkturpaket/20200603-konjunkturpaket-beschlossen.html> [Last retrieved: 13.08.2020].
- 
- [14] Konrad-Adenauer-Stiftung (Ed.), **Vergleich nationaler Strategien zur Förderung von Künstlicher Intelligenz**. Sankt Augustin/Berlin, 2018.
- 
- [15] **COM/2020/65, White Paper on Artificial Intelligence: a European approach to excellence and trust**.
- 
- [16] S. Fouse, S. Cross and Z. Lapin, **DARPA's Impact on Artificial Intelligence**. AI Magazine, 41(2), pp. 3-8, Summer 2020.
-

- 
- [17] Ethik-Kommission, **Automatisiertes und Vernetztes Fahren**. Berlin: BMVI, 2017. [Online]. Available at: [https://www.bmvi.de/SharedDocs/DE/Publikationen/DG/bericht-der-ethik-kommission.pdf?\\_\\_blob=publicationFile](https://www.bmvi.de/SharedDocs/DE/Publikationen/DG/bericht-der-ethik-kommission.pdf?__blob=publicationFile) [Last retrieved: 01.07.2020].
- 
- [18] Deutscher Bundestag, **Enquete-Kommission “Künstliche Intelligenz – Gesellschaftliche Verantwortung und wirtschaftliche, soziale und ökologische Potenziale”**. [Online]. Available at: [https://www.bundestag.de/ausschuesse/weitere\\_gremien/enquete\\_ki](https://www.bundestag.de/ausschuesse/weitere_gremien/enquete_ki) [Last retrieved: 01.07.2020].
- 
- [19] Deutscher Bundestag, **Enquete-Kommission KI, “KI und Wirtschaft”**. Kommissionsdrucksache 19(27)92, 19.12.2019.
- 
- [20] Deutscher Bundestag, **Enquete-Kommission KI, “KI und Staat”**. Kommissionsdrucksache 19(27)93, 19.12.2019.
- 
- [21] Deutscher Bundestag, **Enquete-Kommission KI, “KI und Gesundheit”**. Kommissionsdrucksache 19(27)94, 19.12.2019.
- 
- [22] AI HLEG, **Policy and investment recommendations for trustworthy Artificial Intelligence**. Brussels: European Commission, 2019 [Online]. Available at: <https://ec.europa.eu/digital-single-market/en/news/policy-and-investment-recommendations-trustworthy-artificial-intelligence> [Last retrieved: 13.08.2020].
- 
- [23] PI4.0, **Technologieszenario “Künstliche Intelligenz in der Industrie 4.0”**. Berlin: BMWi, 2019. [Online]. Available at: <https://www.plattform-i40.de/PI40/Redaktion/DE/Downloads/Publikation/KI-industrie-40.html> [Last retrieved 07.08.2020].
- 
- [24] PI4.0, **KI in der Industrie 4.0: Orientierung, Anwendungsbeispiele, Handlungsempfehlungen**. Berlin: BMWi, 2019. [Online]. Available at: <https://www.plattform-i40.de/PI40/Redaktion/DE/Downloads/Publikation/ki-in-der-industrie-40-orientierung-anwendungsbeispiele-handlungsempfehlungen.html> [Last retrieved 07.08.2020].
- 
- [25] PI4.0, **Details of the Asset Administration Shell Part 1 – The exchange of information between partners in the value chain of Industrie 4.0 (Version 2.0.1)**. Berlin: BMWi, 2019. [Online]. Available at: <https://www.plattform-i40.de/PI40/Redaktion/DE/Downloads/Publikation/Details-of-the-Asset-Administration-Shell-Part1.html> [Last retrieved 07.08.2020].
- 
- [26] PI4.0, **Umgang mit Sicherheitsrisiken industrieller Anwendungen durch mangelnde Erklärbarkeit von KI-Ergebnissen**. Berlin: BMWi, 2019. [Online]. Available at: <https://www.plattform-i40.de/PI40/Redaktion/DE/Downloads/Publikation/Umgang-mit-Sicherheitsrisiken.html> [Last retrieved 07.08.2020].
- 
- [27] PI4.0, **Künstliche Intelligenz in Sicherheitsaspekten der Industrie 4.0**. Berlin: BMWi, 2019. [Online]. Available at: <https://www.plattform-i40.de/PI40/Redaktion/DE/Downloads/Publikation/KI-in-sicherheitsaspekten.html> [Last retrieved 07.08.2020].
- 
- [28] PI4.0, **Künstliche Intelligenz und Recht im Kontext von Industrie 4.0**. Berlin: BMWi, 2019. [Online]. Available at: <https://www.plattform-i40.de/PI40/Redaktion/DE/Downloads/Publikation/kuenstliche-intelligenz-und-recht.html> [Last retrieved 07.08.2020].
- 
- [29] PI4.0, **KI und Robotik im Dienste der Menschen**. Berlin: BMWi, 2019. [Online]. Available at: <https://www.plattform-i40.de/PI40/Redaktion/DE/Downloads/Publikation/BMWi%20KI%20und%20Robotik.html> [Last retrieved 07.08.2020].
- 
- [30] AI HLEG, **A definition of AI: Main capabilities and scientific disciplines**. Brussels: European Commission, 2020 [Online]. Available at: [https://ec.europa.eu/newsroom/dae/document.cfm?doc\\_id=56341](https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=56341) [Last retrieved: 10.06.2020].
- 
- [31] S. Russell, P. Norvig, **Artificial Intelligence: A Modern Approach**. 3rd ed. Harrow, UK: Pearson, 2016.
- 
- [32] OECD, **Recommendation of the Council on Artificial Intelligence**. 2019. [Online]. Available at: <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL0449> [Last retrieved: 10.06.2020].
-

- 
- [33] W. Wahlster, **“Künstliche Intelligenz als Grundlage autonomer Systeme”**. Informatik-Spektrum, 40, pp. 409-418, 2017.
- 
- [34] ISO/IEC 20546:2019, **Information technology – Big data – Overview and vocabulary**.
- 
- [35] Auto Zeitung, 4/2019.
- 
- [36] W. Hildesheim, T. Holoyad, T. Schmidt, K. Schuhmacher, **Managing and Understanding Artificial Intelligence Solutions – The AI-Methods, Capabilities and Criticality Grid and its Value for Decision Makers, Developers and Regulators**. Beuth, 1st edition, 2020.
- 
- [37] D. R. Krathwohl, **A Revision of Bloom’s Taxonomy**. Theory into Practice, 41(4), pp. 212-218, 2002.
- 
- [38] DIN EN ISO/IEC 17000:2019-05, **Conformity assessment – Vocabulary and general principles (ISO/IEC 17000:2020); Trilingual version EN ISO/IEC 17000:2020**.
- 
- [39] DIN EN ISO/IEC 17020:2012-07, **Conformity assessment – Requirements for the operation of various types of bodies performing inspection (ISO/IEC 17020:2012)**.
- 
- [40] DIN EN ISO/IEC 17021-1:2015-11, **Conformity assessment – Requirements for bodies providing audit and certification of management systems – Part 1: Requirements (ISO/IEC 17021-1:2015)**.
- 
- [41] DIN EN ISO/IEC 17024:2012-11, **Conformity assessment – General requirements for bodies operating certification of persons (ISO/IEC 17024:2012)**.
- 
- [42] DIN EN ISO/IEC 17025:2018-03, **General requirements for the competence of testing and calibration laboratories (ISO/IEC 17025:2017)**.
- 
- [43] DIN EN ISO/IEC 17029:2020-02, **Conformity Assessment – General principles and requirements for validation and verification bodies (ISO/IEC 17029:2019)**.
- 
- [44] DIN EN ISO/IEC 17065:2013-01, **Conformity assessment – Requirements for bodies certifying products, processes and services (ISO/IEC 17065:2012)**.
- 
- [45] M. Poretschkin, F. Rostalski, J. Voosholz et al., **Vertrauenswürdiger Einsatz von Künstlicher Intelligenz**. Fraunhofer IAIS, 2019. [Online]. Available at: [https://www.iais.fraunhofer.de/content/dam/iais/KINRW/Whitepaper\\_KI-Zertifizierung.pdf](https://www.iais.fraunhofer.de/content/dam/iais/KINRW/Whitepaper_KI-Zertifizierung.pdf) [Last retrieved: 10.08.2020].
- 
- [46] ISO, **WTO ISO Standards Information Gateway**. 2020. [Online]. Available at: <https://tbtcode.iso.org/sites/wto-tbt/home.html> [Last retrieved: 01.07.2020].
- 
- [47] CC 3.1:2017, **Common Criteria for Information Technology Security Evaluation**. [Online]. Available at: <https://www.commoncriteriaportal.org/cc/> [last retrieved: 10.08.2020]
- 
- [48] ISO/IEC 15408-1:2009, **Information technology – Security techniques – Evaluation criteria for IT security – Part 1: Introduction and general model**.
- 
- [49] ISO/IEC 15408-2:2008, **Information technology – Security techniques – Evaluation criteria for IT security – Part 2: Security functional components**.
- 
- [50] ISO/IEC 15408-3:2008, **Information technology – Security techniques – Evaluation criteria for IT security – Part 3: Security assurance components**.
-

- 
- [51] ISO/IEC 1804-5:2008, **Information technology – Security techniques – Methodology for IT security evaluation.**
- 
- [52] CCRA, **Arrangement on the Recognition of Common Criteria Certificates in the field of Information Technology Security.** 2020. [Online]. Available at: <https://www.commoncriteriaportal.org/ccra/> [Last retrieved: 10.08.2020].
- 
- [53] BSI, **Gemeinsame Kriterien für die Prüfung und Bewertung der Sicherheit von Informationstechnik.** 2020. [Online]. Available at: [https://www.bsi.bund.de/DE/Themen/ZertifizierungundAnerkennung/Produktzertifizierung/ZertifizierungnachCC/ITSicherheitskriterien/CommonCriteria/commoncriteria\\_node.html](https://www.bsi.bund.de/DE/Themen/ZertifizierungundAnerkennung/Produktzertifizierung/ZertifizierungnachCC/ITSicherheitskriterien/CommonCriteria/commoncriteria_node.html) [Last retrieved: 10.08.2020].
- 
- [54] ISO/IEC 38500:2015, **Information technology – Governance of IT for the organization.**
- 
- [55] ISO, IEC, ISO/IEC Directives Part 1: Consolidated ISO Supplement, Procedures Specific to ISO, 2020. [Online]. Available at: <https://www.iso.org/sites/directives/current/consolidated/index.xhtml> [Last retrieved: 01.07.2020]. [Anhang L].
- 
- [56] Informationstechnikzentrum Bund, **V-Modell XT: Vorgehensmodell zum Planen und Durchführen von Systementwicklungs-Projekten.** Bonn, 2020. [Online]. Available at: [https://www.itzbund.de/DE/Produkte/V-Modell-XT/vmodell-xt\\_node.html](https://www.itzbund.de/DE/Produkte/V-Modell-XT/vmodell-xt_node.html) [Last retrieved: 01.07.2020].
- 
- [57] Beauftragte der Bundesregierung für Informationstechnik, **Das V-Modell XT.** Berlin: BMI, 2020. [Online]. Available at: [https://www.cio.bund.de/Web/DE/Architekturen-und-Standards/V-Modell-XT/vmodell\\_xt\\_node.html](https://www.cio.bund.de/Web/DE/Architekturen-und-Standards/V-Modell-XT/vmodell_xt_node.html) [Last retrieved: 01.07.2020].
- 
- [58] ISO/IEC/IEEE 12207:2008-02, **Systems and software engineering – Software life cycle processes.**
- 
- [59] ISO 26262-1:2018, **Road vehicles – Functional safety – Part 1: Vocabulary.**
- 
- [60] ISO 26262-2:2018, **Road vehicles – Functional safety – Part 2: Management of functional safety.**
- 
- [61] ISO 26262-3:2018, **Road vehicles – Functional safety – Part 3: Concept phase.**
- 
- [62] ISO 26262-4:2018, **Road vehicles – Functional safety – Part 4: Product development at the system level.**
- 
- [63] ISO 26262-5:2018, **Road vehicles – Functional safety – Part 5: Product development at the hardware level.**
- 
- [64] ISO 26262-6:2018, **Road vehicles – Functional safety – Part 6: Product development at the software level.**
- 
- [65] ISO 26262-7:2018, **Road vehicles – Functional safety – Part 7: Production, operation, service and decommissioning.**
- 
- [66] ISO 26262-8:2018, **Road vehicles – Functional safety – Part 8: Supporting processes.**
- 
- [67] ISO 26262-9:2018, **Road vehicles – Functional safety – Part 9: Automotive safety integrity level (ASIL)-oriented and safety-oriented analyses.**
- 
- [68] ISO 26262-10:2018, **Road vehicles – Functional safety – Part 10: Guidelines on ISO 26262.**
- 
- [69] ISO 26262-11:2018, **Road vehicles – Functional safety – Part 11: Guidelines on application of ISO 26262 to semiconductors.**
- 
- [70] ISO 26262-12:2018, **Road vehicles – Functional safety – Part 12: Adaptation of ISO 26262 for motorcycles.**
- 
- [71] ISO/IEC 27034-1:2011, **Information technology – Security techniques – Application security – Part 1: Overview and concepts.**
-

- 
- [72] ISO/IEC 27034-1:2011/Cor 1:2014, **Information technology – Security techniques – Application security – Part 1: Overview and concepts – Technical Corrigendum 1.**
- 
- [73] ISO/IEC 27034-2:2015, **Information technology – Security techniques – Application security – Part 2: Organization normative framework.**
- 
- [74] ISO/IEC 27034-3:2018, **Information technology – Application security – Part 3: Application security management process.**
- 
- [75] ISO/IEC 27034-5:2017, **Information technology – Security techniques – Application security – Part 5: Protocols and application security controls data structure.**
- 
- [76] ISO/IEC TS 27034-5-1:2018, **Information technology – Application security – Part 5-1: Protocols and application security controls data structure, XML schemas.**
- 
- [77] ISO/IEC 27034-6:2016, **Information technology – Security techniques – Application security – Part 6: Case studies.**
- 
- [78] ISO/IEC 27034-7:2018, **Information technology – Application security – Part 7: Assurance prediction framework.**
- 
- [79] DIN EN 61508 Supplement 1:2005-10; VDE 0803 Supplement 1:2005-10, **Functional safety of electrical/electronic/programmable electronic safety-related systems – Part 0: Functional safety and IEC 61508 (IEC/TR 61508-0:2005).**
- 
- [80] DIN EN 61508-1:2011-02; VDE 0803-1:2011-02, **Functional safety of electrical/electronic/programmable electronic safety-related systems – Part 1: General requirements (IEC 61508-1:2010).**
- 
- [81] DIN EN 61508-2:2011-02; 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).**
- 
- [82] DIN EN 61508-3:2011-02; VDE 0803-3:2011-02, **Functional safety of electrical/electronic/programmable electronic safety-related systems – Part 3: Software requirements (IEC 61508-3:2010).**
- 
- [83] DIN EN 61508-4:2011-02; VDE 0803-4:2011-02, **Functional safety of electrical/electronic/programmable electronic safety-related systems – Part 4: Definitions and abbreviations (IEC 61508-4:2010).**
- 
- [84] 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).**
- 
- [85] DIN EN 61508-6:2011-02; VDE 0803-6:2011-02, **Functional safety of electrical/electronic/programmable electronic safety-related systems – Part 6: Guidelines on the application of IEC 61508-2 and IEC 61508-3 (IEC 61508-6:2010).**
- 
- [86] DIN EN 61508-7:2011-02; VDE 0803-7:2011-02, **Functional safety of electrical/electronic/programmable electronic safety-related systems – Part 7: Overview of techniques and measures (IEC 61508-7:2010).**
- 
- [87] DIN SPEC 92001-1:2019-04, **Artificial Intelligence – Life Cycle Processes and Quality Requirements – Part 1: Quality Meta Model.**
- 
- [88] DKE, **Referenzmodell für eine vertrauenswürdige KI: Erarbeitung einer neuen VDE-Anwendungsregel**, Frankfurt a. M., 2020. [Online]. Available at: <https://www.dke.de/de/news/2019/referenzmodell-vertrauenswuerdige-ki-vde-anwendungsregel> [Last retrieved: 01.07.2020].
-



- 
- [89] ISO/IEC 25012:2008-12, **Software engineering – Software product Quality Requirements and Evaluation (SQuaRE) – Data quality model**.
- 
- [90] L. Bruns, B. Dittwald, F. Meiners et al., **Leitfaden für qualitativ hochwertige Daten und Metadaten**. Berlin: Fraunhofer FOKUS, 2019. [Online]. Available at: [https://cdn0.scrvt.com/fokus/551bf951bf1982f5/0c96bf464ef/NQDM\\_Leitfaden\\_2019.pdf](https://cdn0.scrvt.com/fokus/551bf951bf1982f5/0c96bf464ef/NQDM_Leitfaden_2019.pdf) [Last retrieved: 01.07.2020].
- 
- [91] ISO 8601(2 Teile):2019-02, **Date and time – Representations for information interchange**.
- 
- [92] D. Keim, K.U. Sattler, K., **Von Daten zu KI. Intelligentes Datenmanagement als Basis für Data Science und den Einsatz Lernender Systeme**. München, 2020 [Online]. Available at: [https://www.plattform-lernende-systeme.de/files/Downloads/Publikationen/AG1\\_Whitepaper\\_Von\\_Daten\\_zu\\_KI.pdf](https://www.plattform-lernende-systeme.de/files/Downloads/Publikationen/AG1_Whitepaper_Von_Daten_zu_KI.pdf) [Last retrieved: 30.10.2020].
- 
- [93] DIN ISO 31000:2018-10, **Risk management – Guidelines (ISO 31000:2018)**.
- 
- [94] **Directive 2006/42/EC** of the European Parliament and of the Council of 17 May 2006 on machinery and amending Directive 95/16/EC (recast).
- 
- [95] **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).
- 
- [96] J. Angwin et al., **Machine Bias: There’s Software Used Across the Country to Predict Future Criminals. And It’s Biased Against Blacks**. ProPublica, 23.03.2016. [Online]. Available at: <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing> [Last retrieved: 28.07.2020].
- 
- [97] ACLU, **Smart Reform is Possible**. New York, 2011. [Online]. Available at: <https://www.aclu.org/files/assets/smartreformispossible.pdf> [Last retrieved: 01.07.2020].
- 
- [98] CivilRights.org, **More than 100 Civil Rights, Digital Justice, and Community-Based Organizations Raise Concerns About Pretrial Risk Assessment. The Leadership Conference on Civil and Human Rights und The Leadership Conference Education Fund**. 2018. [Online]. Available at: <https://civilrights.org/2018/07/30/more-than100-civil-rights-digital-justice-and-community-based-organizations-raise-concerns-about-pretrial-risk-assessment/> [Last retrieved: 01.07.2020].
- 
- [99] D. Collingridge, **The social control of technology**. New York: St. Martin’s Press, 1980.
- 
- [100] D. Watson, **The Rhetoric and Reality of Anthropomorphism in Artificial Intelligence**. Minds & Machines, 29, pp. 417-440, 2019.
- 
- [101] E. Brethenoux, **An ode to the analytics grease monkeys**. KDnuggets, 2017. [Online]. Available at: <https://www.kdnuggets.com/2017/02/analytics-grease-monkeys.html> . [Last retrieved: 01.07.2020].
- 
- [102] R. Berk et al. **Fairness in Criminal Justice Risk Assessments: The State of the Art**. Sociological Methods & Research, doi: 10.1177/0049124118782533, 2018. [p. 33].
- 
- [103] B. Lepri et al. **Fair, Transparent, and Accountable Algorithmic Decision-Making Processes: The Premise, the Proposed Solutions, and the Open Challenges**. *Philosophy & Technology* 31, 4, pp. 611–27, 2018. [Here p. 624].
- 
- [104] M. Veale , M. Van Kleek, R. Binns ,**Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making**. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems – CHI ’18, Montreal QC, Canada*, 2018, pp. 1–14. ACM Press. doi: 10.1145/3173574.3174014.
-

- [105] DIN EN ISO 9000:2015-11, **Quality management systems – Fundamentals and vocabulary (ISO 9000:2015)**.
- [106] K. A. Zweig et al., **Wo Maschinen irren können. Verantwortlichkeiten und Fehlerquellen in Prozessen algorithmischer Entscheidungsfindung**. Gütersloh: Bertelsmann Stiftung, 2018. [Online]. Available at: <https://www.bertelsmann-stiftung.de/fileadmin/files/BSt/Publikationen/GrauePublikationen/WoMaschinenIrrenKoennen.pdf> [Last retrieved: 14.07.2020].
- [107] J. Walker, **Meet the New Boss: Big Data. Companies Trade In Hunch-Based Hiring for Computer Modeling**. The Wall Street Journal, 20.09.2012.
- [108] S. Barocas, A. D. Selbst, **Big data's disparate impact**. California Law Review, 104, pp. 671–732, 2016.
- [109] S. Beck et al., **Künstliche Intelligenz und Diskriminierung – Herausforderungen und Lösungsansätze**. München: PLS, 2019. [Online]. Available at: [https://www.plattform-lernende-systeme.de/files/Downloads/Publikationen/AG3\\_Whitepaper\\_250619.pdf](https://www.plattform-lernende-systeme.de/files/Downloads/Publikationen/AG3_Whitepaper_250619.pdf) [Last retrieved: 28.07.2020].
- [110] G. S. Leventhal, J. Karuza, W. R. Fry, **Beyond fairness: A theory of allocation preferences – Justice and social interaction**. 3(1), pp. 167-218, 1980.
- [111] L. Floridi et al., **AI4People: An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations**. Minds and Machines, 28, pp. 689–707, 2018.
- [112] Beijing Academy of Artificial Intelligence (BAAI), **Beijing AI Principles**. 2019. [Online]. Available at: <https://www.baai.ac.cn/blog/beijing-ai-principles> [Last retrieved: 03.03.2020].
- [113] U. Garzcarek, D. Steuer, **Approaching Ethical Guidelines for Data Scientists**, in: N. Bauer et al. (Ed.), **Applications in Statistical Computing: From Music Data Analysis to Industrial Quality Improvement**. Cham: Springer, 2019, pp. 151-169.
- [114] J. Fjeld et al., **Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI**. Berkman Klein Center for Internet & Society, 2020.
- [115] A. Jobin et al., **The global landscape of AI ethics guidelines**. Nature Machine Intelligence, 1, 9, pp. 389–399, 2019.
- [116] T. Hagendorff, **The Ethics of AI Ethics: An Evaluation of Guidelines**. Minds & Machines, 30, 1, pp. 99120, 2020.
- [117] J. Heesen et al., **Ethik-Briefing. Leitfaden für eine verantwortungsvolle Entwicklung und Anwendung von KI-Systemen**. München, 2020. [Online]. Available at: [https://www.plattform-lernende-systeme.de/files/Downloads/Publikationen/AG3\\_Whitepaper\\_EB\\_200831.pdf](https://www.plattform-lernende-systeme.de/files/Downloads/Publikationen/AG3_Whitepaper_EB_200831.pdf).
- [118] N. Huchler et al., **Kriterien für die Mensch-Maschine: Interaktion bei KI. Ansätze für die menschengerechte Gestaltung in der Arbeitswelt**. Plattform Lernende Systeme, München, 2020. [Online]. Available at: [https://www.plattform-lernende-systeme.de/files/Downloads/Publikationen/AG2\\_Whitepaper2\\_220620.pdf](https://www.plattform-lernende-systeme.de/files/Downloads/Publikationen/AG2_Whitepaper2_220620.pdf) [Last retrieved: 28.07.2020].
- [119] L. Inhoffen, **Künstliche Intelligenz: Deutsche sehen eher die Risiken als den Nutzen**. YouGov, 11.09.2018. [Online]. Available at: <https://yougov.de/news/2018/09/11/kunstliche-intelligenz-deutsche-sehen-eher-die-ris/> [Last retrieved: 11.08.2020].
- [120] DIN EN ISO 9001:2015-11, **Quality management systems – Requirements (ISO 9001:2015)**.
- [121] ISO/IEC Guide 51:2014-04, **Safety aspects – Guidelines for their inclusion in standards**.
-

- 
- [122] ISO/IEC 27001:2013, **Information technology – Security techniques – Information security management systems – Requirements**.
- 
- [123] S. Hallensleben et al., **From Principles to Practice – An interdisciplinary framework to operationalize AI ethics**. Gütersloh: Bertelsmann Stiftung, 2020. [Online]. Available at: [https://www.bertelsmann-stiftung.de/fileadmin/files/BSt/Publikationen/GrauePublikationen/WKIO\\_2020\\_final.pdf](https://www.bertelsmann-stiftung.de/fileadmin/files/BSt/Publikationen/GrauePublikationen/WKIO_2020_final.pdf) [Last retrieved: 28.07.2020].
- 
- [124] DIN EN ISO 12100:2011-03, **Safety of machinery – General principles for design – Risk assessment and risk reduction (ISO 12100:2010)**.
- 
- [125] DIN EN ISO 12100 Correction 1:2013-08, **Safety of machinery – General principles for design – Risk assessment and risk reduction**.
- 
- [126] DIN EN ISO 13849-1:2016-06, **Safety of machinery – Safety-related parts of control systems – Part 1: General principles for design (ISO 13849-1:2020)**.
- 
- [127] DIN EN ISO 13849-2, **Safety of machinery – Safety-related parts of control systems – Part 2: Validation (ISO 13849-2:2012)**.
- 
- [128] DIN EN ISO 14971:2020-07, **Medical devices – Application of risk management to medical devices (ISO 14971:2019)**.
- 
- [129] DIN EN 62061:2016-05; VDE 0113-50:2016-05, **Safety of machinery – Functional safety of safety-related electrical, electronic and programmable electronic control systems (IEC 62061:2005 + A1:2012 + A2:2015)**.
- 
- [130] D. Dawson, E. Schleiger et al., **Artificial Intelligence: Australia’s Ethics Framework**. Data61 CSIRO, Australia, 2019. [Online]. Available at: [https://consult.industry.gov.au/strategic-policy/artificial-intelligence-ethics-framework/supporting\\_documents/ArtificialIntelligenceethicsframeworkdiscussionpaper.pdf](https://consult.industry.gov.au/strategic-policy/artificial-intelligence-ethics-framework/supporting_documents/ArtificialIntelligenceethicsframeworkdiscussionpaper.pdf) [Last retrieved: 14.08.2020].
- 
- [131] J. Black, **Risk-based Regulation: Choices, Practices and Lessons Being Learnt**. in: OECD, **Risk and Regulatory Policy: Improving the Governance of Risk**. Paris, 2010, pp. 185–236.
- 
- [132] I. MacNeil, **Risk control strategies: an assessment in the context of the credit crisis**. in: I. MacNeil, J. O’Brien (Ed.), **The future of financial regulation**. Oxford, Portland: Hart Pub, 2010, pp. 141–160.
- 
- [133] J. Black, R. Baldwin, **When risk-based regulation aims low: a strategic framework**. *Regulation & Governance*, 6(2), pp. 131–148, 2012.
- 
- [134] F. Saurwein et al., **Governance of algorithms: options and limitations**. *info*, 17(6), pp. 35–49, 2015.
- 
- [135] M. Z. van Drunen et al., **Know your algorithm: what media organizations need to explain to their users about news personalization**. *International Data Privacy Law*, 9(4), pp. 220–235, 2019.
- 
- [136] T. D. Krafft, K. A. Zweig, **Transparenz und Nachvollziehbarkeit algorithmenbasierter Entscheidungsprozesse: Ein Regulierungsvorschlag aus sozioinformatischer Perspektive**. Berlin: Verbraucherzentrale Bundesverband e.V., 2019. [Online]. Available at: [https://www.vzbv.de/sites/default/files/downloads/2019/05/02/19-01-22\\_zweig\\_krafft\\_transparenz\\_adm-neu.pdf](https://www.vzbv.de/sites/default/files/downloads/2019/05/02/19-01-22_zweig_krafft_transparenz_adm-neu.pdf) [Last retrieved: 06.08.2020].
- 
- [137] Bundesärztekammer, **(Muster)Berufsordnung für die in Deutschland tätigen Ärztinnen und Ärzte – MBO-Ä 1997 –\* in der Fassung der Beschlüsse des 121. Deutschen Ärztetages 2018 in Frankfurt am Main**. *Deutsches Ärzteblatt*, 01.02.2019, pp. A1-A9. [here S 9, p. A4].
- 
- [138] C. Reed, **How should we regulate artificial intelligence?**. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376, 20170360, 2018.
-

- [139] N. Diakopoulos, **Algorithmic accountability reporting: On the investigation of black boxes**. Tow Center for Digital Journalism Publications, 2014.
- [140] W3C, **Semantic Web**. [Online.] Available at: [https://www.w3.org/2001/sw/wiki/Main\\_Page](https://www.w3.org/2001/sw/wiki/Main_Page) [Last retrieved: 11.08.2020].
- [141] **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.
- [142] **Bundesdatenschutzgesetz (Federal Data Protection Act) (BDSG)**.
- [143] DSK, **Hambacher Erklärung zur Künstlichen Intelligenz**. Entschließung der 97. Konferenz der unabhängigen Datenschutzaufsichtsbehörden des Bundes und der Länder, Hambacher Schloss, 3. April 2019. [Online]. Available at: [https://www.datenschutzkonferenz-online.de/media/en/20190405\\_hambacher\\_erklaerung.pdf](https://www.datenschutzkonferenz-online.de/media/en/20190405_hambacher_erklaerung.pdf) [Last retrieved: 11.08.2020].
- [144] **JCGM 200:2012, International vocabulary of metrology – Basic and general concepts and associated terms (VIM) [= ISO/IEC Guide 99]**.
- [145] DIN EN ISO 19011:2018-10, **Guidelines for auditing management systems (ISO 19011:2018)**.
- [146] ISO/IEC 25010:2011, **Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – System and software quality models**.
- [147] S. Goericke (ed.), **The Future of Software Quality Assurance**. Basel: Springer, 2020.
- [148] ISO/PAS 21448:2019, **Road vehicles – Safety of the intended functionality**.
- [149] ISO/IEC TR 20547-2:2018, **Information technology – Big data reference architecture – Part 2: Use cases and derived requirements**.
- [150] ISO/IEC TR 20547-5:2018, **Information technology – Big data reference architecture – Part 5: Standards roadmap**.
- [151] DIN SPEC 13266:2020-04, **Guideline for the development of deep learning image recognition systems**.
- [152] **ETSI TS 103 296 V1.1.1:2016-08, Speech and Multimedia Transmission Quality (STQ) – Requirements for Emotion Detectors used for Telecommunication Measurement Applications – Detectors for written text and spoken speech**.
- [153] **ETSI TS 103 1952 V1.1.1:2018-05, 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**.
- [154] **ITU-T Y.3170:2018-09, Requirements for machine learning-based quality of service assurance for the IMT-2020 network**.
- [155] **ITU-T Y.3173:2020-02, Framework for evaluating intelligence levels of future networks including IMT-2020**.
- [156] **IEEE 7010:2020, IEEE Recommended Practice for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-Being**.
- [157] **UL 4600:2020-04, Evaluation of Autonomous Products**.
-

- 
- [158] T. Weinberger et al., **Modelle Maschinellen Lernens: Symbolische und konnektionistische Ansätze**. Karlsruhe: Kernforschungszentrum Karlsruhe, 1994. [= KfK 5184].
- 
- [159] G. Kern-Isberner, **Fortgeschrittene Themen der Wissensrepräsentation**. Vorlesung TU Dortmund, 2020. [Online.] Zusammenfassung verfügbar unter: <https://ls1-www.cs.tu-dortmund.de/de/lehveranstaltungen/503-ftw-ss-2020/1897-fortgeschrittene-themen-der-wissensrepr%C3%A4sentation-ss-20> [Last retrieved: 11.08.2020].
- 
- [160] C. E. Alchourròn et al., **On the logic of theory change: Partial meet contraction and revision functions**. *Journal of Symbolic Logic*, 50(2), pp. 510–530, 1985.
- 
- [161] D. Mackall et al., **Verification and validation of neural networks for aerospace systems**. Mofett Field (CA): NASA, 2002.
- 
- [162] N. Röttger et al., **Warum KI auch eine intelligente Qualitätssicherung braucht**. *OBJEKTSpektrum*, 02/2020, pp. 20-24.
- 
- [163] ISO/IEC 27701:2019, **Security techniques – Extension to ISO/IEC 27001 and ISO/IEC 27002 for privacy information management – Requirements and guidelines**.
- 
- [164] IATF 16949:2016, **Quality management system requirements for automotive production and relevant service parts organisations**.
- 
- [165] DIN EN 9100:2018-08, **Quality Management Systems – Requirements for Aviation, Space and Defence Organizations**.
- 
- [166] H. W. Dörmann Osuna, **Ansatz für ein prozessintegriertes Qualitätsregelungssystem für nicht-stabile Prozesse**. Dissertation, Techn. Univ. Ilmenau, 2008.
- 
- [167] Hering et al, Qualitätslenkung mit Produkt-Regelkreis, in: E. Hering, J. Triemel, H.-P. Blank (Ed.), **Qualitätsmanagement für Ingenieure**. Berlin, Heidelberg: Springer, 2003, Chapter E3.3.
- 
- [168] R. Schmitt, T. Pfeifer, **Qualitätsmanagement: Strategien, Methoden, Techniken**. München, Wien: Carl Hanser Verlag, 2015.
- 
- [169] G. Montavon et al., **Methods for Interpreting and Understanding Deep Neural Networks**. arXiv:1706.07979, 2017.
- 
- [170] E. Štrumbelj, I. Kononenko, **Explaining prediction models and individual predictions with feature contributions**. *Knowledge and Information Systems*, 41, pp. 647–665, 2014.
- 
- [171] A. Shrikumar et al., **Learning Important Features Through Propagating Activation Differences**. arXiv:1704.02685, 2019.
- 
- [172] A. Datta et al., **Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems**. 2016 IEEE Symposium on Security and Privacy (SP), San Jose, CA, 2016, pp. 598-617.
- 
- [173] T. Menzies, **Verification and Validation and Artificial Intelligence: Beyond the State of the Art**. Foundations 02: A V&V Workshop, Johns Hopkins University Laurel, Maryland USA, 2002.
- 
- [174] R. Ehlers, **Formal verification of piece-wise linear feed-forward neural networks**. arXiv:1705.01320v3, 2017.
- 
- [175] **ETSI TR 101 583 V 1.1.1:201503, Methods for Testing and Specification (MTS) – Security Testing – Basic Terminology**.
- 
- [176] A. Odena, I. Goodfellow, **TensorFuzz: Debugging Neural Networks with Coverage-Guided Fuzzing**. arXiv:1807.10875, 2018.
-

- [177] J. Guo, et al., **Dlfuzz: Differential fuzzing testing of deep learning systems**. Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, pp. 739-743, 2018.
- [178] **NISTIR 8269, A Taxonomy and Terminology of Adversarial Machine Learning (draft)**.
- [179] Y. Dong et al., **There is Limited Correlation between Coverage and Robustness for Deep Neural Networks**. arXiv:1911.05904, 2019.
- [180] Spiegel Netzwelt, 10.06.2020. [Online]. Available at: <https://www.spiegel.de/netzwelt/web/honda-muss-produktion-nach-cyberangriff-stoppen-a-69f59803-216c-43c7-9ee9-59d3926d6314#> [Last retrieved: 11.08.2020].
- [181] Welt, 19.06.2020. [Online.] Available at: <https://www.welt.de/politik/ausland/article209883177/Cyber-Angriffe-auf-Australien-Premier-Scott-Morrison-hat-eine-Vermutung.html> [Last retrieved: 11.08.2020].
- [182] A. Berg, M. Niemeier, **Wirtschaftsschutz in der digitalen Welt**. Berlin: Bitkom, 6.11.2019. [Online]. Available at: [https://www.bitkom.org/sites/default/files/2019-11/bitkom\\_wirtschaftsschutz\\_2019.pdf](https://www.bitkom.org/sites/default/files/2019-11/bitkom_wirtschaftsschutz_2019.pdf) [Last retrieved: 17.08.2020].
- [183] Kompass Informationssicherheit und Datenschutz, **IT-Sicherheits- und Risikomanagement**. Berlin: Bitkom, DIN, 2020. [Online]. Available at: <https://www.kompass-sicherheitsstandards.de/ISMS/Allgemeine-ISMS> [Last retrieved: 17.08.2020].
- [184] DIN, DKE, **Deutsche Normungs-Roadmap IT-Sicherheit (German Standardization Roadmap on IT Security)**. Berlin, Frankfurt (M.), Version 3.0, 2017 [Online]. Available at: <https://www.din.de/de/din-und-seine-partner/presse/mitteilungen/normungs-roadmap-it-sicherheit-aktualisiert-238508> [Last retrieved: 17.08.2020].
- [185] **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.
- [186] **Directive (EU) 2016/680** 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 by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data, and repealing Council Framework Decision 2008/977/JHA.
- [187] **Directive 2002/58/EC** of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications).
- [188] **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).
- [189] **Directive 2001/95/EC** of the European Parliament and of the Council of 3 December 2001 on general product safety.
- [190] **Gesetz zur Erhöhung der Sicherheit informationstechnischer Systeme. (IT-Sicherheitsgesetz)**.
- [191] **Zweites Gesetz zur Anpassung des Datenschutzrechts an die Verordnung (EU) 2016/679 und zur Umsetzung der Richtlinie (EU) 2016/680** (Zweites Datenschutz-Anpassungs- und Umsetzungsgesetz EU – 2. DSAnpUG-EU).
- [192] **Telemediengesetz (Telemedia Act) (TMG)**.
- [193] **Telekommunikationsgesetz (Telecommunications Act) (TKG)**.
-

- 
- [194] **Gesetz über das Bundesamt für Sicherheit in der Informationstechnik** (BSI-Gesetz – [BSIG](#)).
- 
- [195] **Gesetz über die Bereitstellung von Produkten auf dem Markt** (Produktsicherheitsgesetz – [ProdSG](#)).
- 
- [196] **Gesetz über die Elektrizitäts- und Gasversorgung** (Energiewirtschaftsgesetz – [EnWG](#)).
- 
- [197] bitkom, **Regulierungsmapping IT-Sicherheit**. Berlin, 08.2019. [Online.] Available at: [https://www.bitkom.org/sites/default/files/2019-08/190816\\_regulierungsmapping.pdf](https://www.bitkom.org/sites/default/files/2019-08/190816_regulierungsmapping.pdf) [Last retrieved: 17.08.2020].
- 
- [198] IEC TS 62443-1-1:2009, **Industrial communication networks – Network and system security – Part 1-1: Terminology, concepts and models**.
- 
- [199] IEC 62443-2-1:2010, **Industrial communication networks – Network and system security – Part 2-1: Establishing an industrial automation and control system security program**.
- 
- [200] IEC TR 62443-2-3:2015, **Security for industrial automation and control systems – Part 2-3: Patch management in the IACS environment**.
- 
- [201] DIN EN IEC 62443-2-4:2020-07; VDE 0802-2-4:2020-07, **Security for industrial automation and control systems – Part 2-4: Part 2-4: Security program requirements for IACS service providers (IEC 62443-2-4:2015 + Cor.:2015 + A1:2017)**.
- 
- [202] IEC TR 62443-3-1:2009, **Industrial communication networks – Network and system security – Part 3-1: Security technologies for industrial automation and control systems**.
- 
- [203] IEC 62443-3-2:2020, **Security for industrial automation and control systems – Part 3-2: Security risk assessment for system design**.
- 
- [204] IEC 62443-3-3:2013, **Industrial communication networks – Network and system security – Part 3-3: System security requirements and security levels**.
- 
- [205] DIN EN IEC 62443-3-3:2020-01; VDE 0802-3-3:2020-01, **Industrial communication networks – Network and system security – Part 3-3: System security requirements and security levels (IEC 62443-3-3:2013 + COR1:2014)**.
- 
- [206] IEC 62443-4-1:2018, **Security for industrial automation and control systems – Part 4-1: Secure product development lifecycle requirements**.
- 
- [207] DIN EN IEC 62443-4-1:2018-10; VDE 0802-4-1:2018-10, **Security for industrial automation and control systems – Part 4-1: Secure product development life cycle requirements (IEC 62443-4-1:2018)**.
- 
- [208] IEC 62443-4-2:2019, **Security for industrial automation and control systems – Part 4-2: Technical security requirements for IACS components**.
- 
- [209] DIN EN IEC 62443-4-2:2019-12; VDE 0802-4-2:2019-12, **Security for industrial automation and control systems – Part 4-2: Technical security requirements for IACS components (IEC 62443-4-2:2019)**.
- 
- [210] ISO/IEC 27005:2018, **Information technology – Security techniques – Information security risk management**.
- 
- [211] 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)**.
- 
- [212] DIN EN ISO/IEC 29134:2020-09, **Information technology – Security techniques – Guidelines for privacy impact assessment (ISO/IEC 29134:2017)**.
-

- [213] Hirsch-Kreinsen et al., **Themenfelder Industrie 4.0: Forschungs- und Entwicklungsbedarfe zur erfolgreichen Umsetzung von Industrie 4.0**. Forschungsbeirat der PI4.0, 2019. [Online]. Available at: [https://www.acatech.de/wp-content/uploads/2019/09/Forschungsbeirat\\_Themenfelder-Industrie-4.0-2.pdf](https://www.acatech.de/wp-content/uploads/2019/09/Forschungsbeirat_Themenfelder-Industrie-4.0-2.pdf) [Last retrieved: 12.08.2020].
- [214] J. Lee, **Industrial AI. Singapore**: Springer, 2020.
- [215] GMA, **VDI-Statusreport Industrie 4.0 Wertschöpfungsketten**. Düsseldorf, 2014. <https://www.vdi.de/ueber-uns/presse/publikationen/details/industrie-40-wertschoepfungsketten>.
- [216] PI4.0, **Fortschreibung der Anwendungsszenarien der Plattform Industrie 4.0**. Berlin: BMWi, 2016. [Online]. Available at: <https://www.plattform-i40.de/PI40/Redaktion/DE/Downloads/Publikation/fortschreibung-anwendungsszenarien.html> [Last retrieved: 12.08.2020].
- [217] SCI4.0, DIN/DKE, **German Standardization Roadmap Industrie 4.0**. Version 4, Berlin: DIN and Frankfurt(M): DKE. [Online]. Available at: <https://www.din.de/de/forschung-und-innovation/themen/industrie40/roadmap-industrie40-62178> [Last retrieved: 12.08.2020].
- [218] E VDE-AR-E 2842-611:2020-07, **Entwicklung und Vertrauenswürdigkeit von autonom/kognitiven Systemen – Teil 611: Terminologie und Grundkonzepte**.
- [219] GMA, **VDI-Statusreport Maschinelles Lernen**. Düsseldorf, 2019.
- [220] VDI/VDE/VDMA 2632 Blatt 2:2015-10, **Machine vision – Guideline for the preparation of a requirement specification and a system specification**.
- [221] VDI/VDE/VDMA 2632 Blatt 3:2017-10, **Machine vision/industrial image processing – Acceptance test of classifying machine vision systems**.
- [222] VDI/VDE/VDMA 2632 Blatt 3.1:2020-08, **Machine vision/industrial image processing – Acceptance test of classifying machine vision systems – Test of classification performance**.
- [223] VDI/VDE/VDMA 2632 Blatt 4.1:2020-08, **Machine vision/industrial image processing – Surface inspection systems in flat steel production – Stability testing**.
- [224] **COM/2018/237, Artificial Intelligence for Europe**.
- [225] D. Schneider, M. Trapp, **Conditional safety certification of open adaptive systems**. ACM Transactions on Autonomous and Adaptive Systems (TAAS), 8, 2013.
- [226] D. Schneider, et al., **WAP: Digital dependability identities**. 2015 IEEE 26th International Symposium on Software Reliability Engineering (ISSRE), Gaithersburg, MD, USA, 2015, pp. 324-329.
- [227] **COM/2020/64, Report on the safety and liability implications of Artificial Intelligence, the Internet of Things and robotics**. Brussels: European Commission.
- [228] **Straßenverkehrsgesetz (StVG)**.
- [229] **Regulation (EU) 2018/858** of the European Parliament and of the Council of 30 May 2018 on the approval and market surveillance of motor vehicles and their trailers, and of systems, components and separate technical units intended for such vehicles, amending Regulations (EC) No 715/2007 and (EC) No 595/2009 and repealing Directive 2007/46/EC.
-



- 
- [230] **Agreement concerning the Adoption of Harmonized Technical United Nations Regulations for Wheeled Vehicles, Equipment and Parts which can be Fitted and/or be Used on Wheeled Vehicles and the Conditions for Reciprocal Recognition of Approvals Granted on the Basis of these United Nations Regulations.** *Official Journal of the European Union*, L 274, pp. 4–30, 11.10.2016.
- 
- [231] **Directive 2014/45/EU** of the European Parliament and of the Council of 3 April 2014 on periodic roadworthiness tests for motor vehicles and their trailers and repealing Directive 2009/40/EC.
- 
- [232] **UN Regulation No. 79.** Uniform provisions concerning the approval of vehicles with regard to steering equipment, Revision 4, 18.10.2018. [Online]. Available at: <https://www.unece.org/fileadmin/DAM/trans/main/wp29/wp29regs/2018/R079r4e.pdf> [Last retrieved: 12.08.2020].
- 
- [233] PLS, **Intelligent vernetzt unterwegs.** [Online]. Available at: <https://www.plattform-lernende-systeme.de/umfeldszenario-intelligent-vernetzt-unterwegs.html> [Last retrieved: 12.08.2020].
- 
- [234] **Straßenverkehrs-Ordnung (StVO).**
- 
- [235] ISO/IEC 11179-1:2015, **Information technology – Metadata registries (MDR) – Part 1: Framework.**
- 
- [236] ISO/IEC TR 11179-2:2019, **Information technology – Metadata registries (MDR) – Part 2: Classification.**
- 
- [237] ISO/IEC 11179-3:2013, **Information technology – Metadata registries (MDR) – Part 3: Registry metamodel and basic attributes.**
- 
- [238] ISO/IEC 11179-4:2004, **Information technology – Metadata registries (MDR) – Part 4: Formulation of data definitions.**
- 
- [239] ISO/IEC 11179-5:2015, **Information technology – Metadata registries (MDR) – Part 5: Naming principles.**
- 
- [240] ISO/IEC 11179-6:2015, **Information technology – Metadata registries (MDR) – Part 6: Registration.**
- 
- [241] ISO/IEC 11179-7:2019, **Information technology – Metadata registries (MDR) – Part 7: Metamodel for data set registration.**
- 
- [242] ISO/IEC TS 11179-30:2019, **Information technology – Metadata registries (MDR) – Part 30: Basic attributes of metadata.**
- 
- [243] ISO/IEC 19763-1:2015, **Information technology – Metamodel framework for interoperability (MFI) – Part 1: Framework.**
- 
- [244] ISO/IEC 19763-3:2010, **Information technology – Metamodel framework for interoperability (MFI) – Part 3: Metamodel for ontology registration.**
- 
- [245] ISO/IEC 19763-5:2015, **Information technology – Metamodel framework for interoperability (MFI) – Part 5: Metamodel for process model registration.**
- 
- [246] ISO/IEC 19763-6:2015, **Information technology – Metamodel framework for interoperability (MFI) – Part 6: Registry Summary.**
- 
- [247] ISO/IEC 19763-7:2015, **Information technology – Metamodel framework for interoperability (MFI) – Part 7: Metamodel for service model registration.**
-

- [248] ISO/IEC 19763-8:2015, **Information technology – Metamodel framework for interoperability (MFI) – Part 8: Metamodel for role and goal model registration.**
- [249] ISO/IEC TR 19763-9:2015, **Information technology – Metamodel framework for interoperability (MFI) – Part 9: On demand model selection.**
- [250] ISO/IEC 19763-10:2014, **Information technology – Metamodel framework for interoperability (MFI) – Part 10: Core model and basic mapping.**
- [251] ISO/IEC 19763-12:2015, **Information technology – Metamodel framework for interoperability (MFI) – Part 12: Metamodel for information model registration.**
- [252] ISO/IEC TS 19763-13:2016, **Information technology – Metamodel framework for interoperability (MFI) – Part 13: Metamodel for form design registration.**
- [253] M. Salathé et al., **Focus Group on Artificial Intelligence for Health.** [Online]. Available at: <https://www.itu.int/en/ITU-T/focusgroups/ai4h/Pages/default.aspx> [Last retrieved: 12.08.2020].
- [254] J. Müller-Quade et al., **Sichere KI-Systeme für die Medizin. Datenmanagement und IT-Sicherheit in der Krebsbehandlung der Zukunft.** München, 2020 [Online] Available at [https://www.plattform-lernende-systeme.de/files/Downloads/Publikationen/AG3\\_6\\_Whitepaper\\_07042020.pdf](https://www.plattform-lernende-systeme.de/files/Downloads/Publikationen/AG3_6_Whitepaper_07042020.pdf) [Last retrieved: 20.10.2020].
- [255] WMA, Declaration of Helsinki: **Ethical Principles for Medical Research Involving Human Subjects.** 2020. [Online]. Available at: <https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects/> [Last retrieved: 10.06.2020].
- [256] **Gesetz über Medizinprodukte** (Medizinproduktegesetz – MPG).
- [257] Acatech, “**Lernende Systeme: Die Plattform für Künstliche Intelligenz**” **Deutsche Akademie der Technikwissenschaften e. V.** 2020. [Online]. Available at: <https://www.plattform-lernende-systeme.de/startseite.html> [Last retrieved: 10.06.2020].
- [258] C. Wischhöfer, P. Rauh, **Standards of the Future – Stand der Arbeiten zum Thema maschinenausführbarer Normen-inhalte, DIN-Mitteilungen,** August 2019, pp. 4-8.
- [259] D. Czarny et al., **Digitale Transformation in der Normung – Ausgangssituation und Vision – Initiative Digitale Standards – Herausforderungen und Ziele.** Seminar I und II am 2.6.2020 und 10.6.2020. [Online]. Verfügbar unter: <https://youtu.be/B4f09mxVHsl> und [https://youtu.be/laEXmB0m\\_PI](https://youtu.be/laEXmB0m_PI) [Last retrieved: 07.08.2020].
- [260] IEC SG12 TF 2, **Digital Transformation (2019) Work Package 1 – Classification Scheme and Use Cases (utility model), IEC.**
- [261] R. Heidel et al., **Industrie 4.0. – The Reference Architecture Model RAMI 4.0 and the Industrie 4.0 Component.** Berlin: Beuth, 2017.
- [262] ISO/IEC TR 24028:2020, **Information technology – Artificial intelligence – Overview of trustworthiness in artificial intelligence.**
- [263] **ETSI GR ENI 004 V 2.1.1:2019-10, Experiential Networked Intelligence (ENI) – Terminology for Main Concepts in ENI.**
- [264] **ETSI GR NFVSEC 003 V 1.2.1:2016-08, Network Functions Virtualisation (NFV) – NFV Security – Security and Trust Guidance.**
-

- 
- [265] ISO/IEC/IEEE 29119-1:2013, **Software and systems engineering – Software testing – Part 1: Concepts and definitions.**
- 
- [266] ISO/IEC/IEEE 29119-2:2013, **Software and systems engineering – Software testing – Part 2: Test processes.**
- 
- [267] ISO/IEC/IEEE 29119-3:2013, **Software and systems engineering – Software testing – Part 3: Test documentation.**
- 
- [268] ISO/IEC/IEEE 29119-4:2015, **Software and systems engineering – Software testing – Part 4: Test techniques.**
- 
- [269] ISO/IEC/IEEE 29119-5:2016, **Software and systems engineering – Software testing – Part 5: Keyword-Driven Testing.**
- 
- [270] ISO/IEC TR 20547-1:2020, **Information technology – Big data reference architecture – Part 1: Framework and application process.**
- 
- [271] ISO/IEC 20547-3:2020, **Information technology – Big data reference architecture – Part 3: Reference architecture.**
- 
- [272] ISO/IEC 25000:2014-03, **Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – Guide to SQuaRE.**
- 
- [273] ISO/IEC 25020:2019-07, **Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – Quality measurement framework.**
- 
- [274] ISO/IEC 25021:2012-11, **Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – Quality measure elements.**
- 
- [275] ISO/IEC 25024:2015-10, **System und Software-Engineering – Qualitätskriterien und Bewertung von System- und Softwareprodukten (SQuaRE) – Messung der Datenqualität.**
- 
- [276] ISO/IEC 27002:2013, **Information technology – Security techniques – Code of practice for information security controls.**
- 
- [277] DIN EN ISO/IEC 29100, **Information technology – Security techniques – Privacy framework (ISO/IEC 29100:2011, including Amd 1:2018).**
- 
- [278] ISO/IEC 33063:2015, **Information technology – Process assessment – Process assessment model for software testing.**
- 
- [279] DIN ISO/TR 22100-1:2016-07; DIN SPEC 33886:2016-07, **Safety of machinery – Relationship with ISO 12100 – Part 1: How ISO 12100 relates to type-B and type-C standards (ISO/TR 22100-1:2015).**
- 
- [280] DIN ISO/TR 22100-2:2014-09; DIN SPEC 33887:2014-09, **Safety of machinery – Relationship with ISO 12100 – Part 2: How ISO 12100 relates to ISO 13849-1 (ISO/TR 22100-2:2013).**
- 
- [281] DIN ISO/TR 22100-3:2017-06; DIN SPEC 33888:2017-06, **Safety of machinery – Relationship with ISO 12100 – Part 3: Implementation of ergonomic principles in safety standards (ISO/TR 22100-3:2016).**
- 
- [282] ISO 23412:2020, **Indirect, temperature-controlled refrigerated delivery services – Land transport of parcels with intermediate transfer.**
- 
- [283] ISO 25119-1:2018-10, **Tractors and machinery for agriculture and forestry – Safety-related parts of control systems – Part 1: General principles for design and development.**
-

- 
- [284] ISO 25119-2:2019-08, **Tractors and machinery for agriculture and forestry – Safety-related parts of control systems – Part 2: Concept phase.**
- 
- [285] ISO 25119-3:2018-10, **Tractors and machinery for agriculture and forestry – Safety-related parts of control systems – Part 3: Series development, hardware and software.**
- 
- [286] ISO 25119-4:2018-10, **Tractors and machinery for agriculture and forestry – Safety-related parts of control systems – Part 4: Production, operation, modification and supporting processes.**
- 
- [287] DIN EN 60601-1-4:2001-04; VDE 0750-1-4:2001-04, **Medical electrical equipment – Part 1-4: General requirements for safety; Collateral standard: Programmable electrical medical systems (IEC 60601-1-4:1996 + A1:1999).**
- 
- [288] DIN EN 61511-2; VDE 0810-2:2019-02, **Functional safety – Safety instrumented systems for the process industry sector – Part 2: Guidelines for the application of IEC 61511-1 (IEC 61511-2:2016).**
- 
- [289] DIN EN 61511-3; VDE 0810-3:2019-02, **Functional safety – Safety instrumented systems for the process industry sector – Part 3: Guidance for the determination of the required safety integrity levels (IEC 61511-3:2016).**
- 
- [290] 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).**
- 
- [291] DIN EN 62304:2016-10; VDE 0750-101:2016-10, **Medical device software – Software life-cycle processes (IEC 62304:2006 + A1:2015).**
- 
- [292] DIN EN 50128:2012-03; VDE 0831-128:2012-03, **Railway applications – Communication, signalling and processing systems – Software for railway control and protection systems.**
- 
- [293] [IEEE 1012:2016](#), **IEEE Standard for System, Software, and Hardware Verification and Validation.**
- 
- [294] ISO/IEC TR 13066-2:2016-02, **Information technology – Interoperability with assistive technology (AT) – Part 2: Windows accessibility application programming interface (API).**
- 
- [295] DIN EN ISO 13482:2014-11, **Robots and robotic devices – Safety requirements for personal care robots (ISO 13482:2014).**
- 
- [296] Wissenschaftsjahr 2019 KI, “**Glossar: Lernende Systeme**”, **Bundesministerium für Bildung und Forschung**. [Online]. Available at: [https://www.wissenschaftsjahr.de/2019/uebergreifende-informationen/glossar/detail/?tx\\_dpnglossary\\_glossarydetail%5Bcontroller%5D=Term&tx\\_dpnglossary\\_glossarydetail%5Baction%5D=show&tx\\_dpnglossary\\_glossarydetail%5Bterm%5D=27&tx\\_dpnglossary\\_glossarydetail%5BpageUid%5D=1016&cHash=2a2f33e3e34125305328e93c376e424a](https://www.wissenschaftsjahr.de/2019/uebergreifende-informationen/glossar/detail/?tx_dpnglossary_glossarydetail%5Bcontroller%5D=Term&tx_dpnglossary_glossarydetail%5Baction%5D=show&tx_dpnglossary_glossarydetail%5Bterm%5D=27&tx_dpnglossary_glossarydetail%5BpageUid%5D=1016&cHash=2a2f33e3e34125305328e93c376e424a) [Last retrieved 12.08.2020].
- 
- [297] ISO/IEC 18023-1:2006-05, **Information technology – SEDRIS language bindings – Part 1: Functional specification.**
- 
- [298] S. Jordan, C Nimtz (Ed.), **Lexikon Philosophie: Hundert Grundbegriffe**. Stuttgart: Reclam, 2009.
- 
- [299] D. Frey, L. K. Schmalzried, **Philosophie der Führung: Gute Führung lernen von Kant, Aristoteles, Popper & Co.** 1st ed. Berlin, Heidelberg: Springer-Verlag, 2013. [p. 62 f.].
- 
- [300] C. Misselhorn, **Grundfragen der Maschinenethik**. 3rd ed. Ditzingen: Reclam, 2018.
- 
- [301] D. von der Pfordten, “Rechtsethik”, in: J. Nina-Rümelin, **Angewandte Ethik: Die Bereichsethiken und ihre theoretische Fundierung: Ein Handbuch**. 2 ed. Stuttgart: Alfred Kröner Verlag, 2005. [pp. 207-208].
-

- 
- [302] M. Lutz-Bachmann, **Grundkurs Philosophie**. Bd. 7.: Ethik. Ditzingen: Reclam, 2013. [p. 201].
- 
- [303] W. Gründinger et al., **Mensch, Moral, Maschine**. Berlin: BVDW, 2019. [Online]. Available at: [https://www.bvdw.org/fileadmin/bvdw/upload/dokumente/BVDW\\_Digitale\\_Ethik.pdf](https://www.bvdw.org/fileadmin/bvdw/upload/dokumente/BVDW_Digitale_Ethik.pdf) [Last retrieved: 12.08.2020]. [p. 20].
- 
- [304] W. Damm, P. Heidl et al. **Roadmap – Safety, Security, and Certifiability Future Man-Machine Systems, SafeTRANS-Arbeitskreises Resilient, Learning, and Evolutionary Systems**. Oldenburg: SafeTRANS, 2019. [Online]. Available at: <https://www.safetrans-de.org/de/Aktuelles/aktuelle-roadmap-%22safety%2C-security%2C-and-certifiability-of-future-man-machine-systems%22/285> [Last retrieved: 12.08.2020].
- 
- [305] C. Wischhöfer, B. Oberbichler, **Normen-Management-Lösungen: Werknormenanalyse durch die DIN Software GmbH, DIN-Mitteilungen**, November 2015, pp. 14.
- 
- [306] M. Esser et al., **Digitale Content-Dienstleistungen aus dem zentralen XML Content Repository: Zentrale Ablage von Inhalten und Trennung der Inhalte von ihrer Darstellungsform. DIN-Mitteilungen**, Oktober 2017, pp. 18-23.
- 
- [307] **ANSI/NISO Z39.102:2017 STS: Standards Tag Suite**.
- 
- [308] NISO, **NISO Working Group to Develop A Standards-Specific Ontology Standard (SSOS)**. [Online]. Available at: <https://www.niso.org/press-releases/2019/02/niso-working-group-develop-standards-specific-ontology-standard-ssos> [Last retrieved 12.08.2020].
- 
- [309] M. Schacht, **SMART Standards – Entwicklungsprozess und Contentstruktur. DIN-Mitteilungen**, Juni 2020, pp. 36-42.
- 
- [310] Ehring, D.; Loibl, A.; Nagarajah, A.; Zhou, L. (2020) **Smart Standards: Automatisierungsansatz – Methodik zur Wissensrepräsentation**.
- 
- [311] A. Loibl et al., **Procedure for the transfer of standards into machine-actionability**. Journal of Advanced Mechanical Design Systems and Manufacturing, 14, JAMDSM0022, 2020.
- 
- [312] VDI 2221 Blatt 1:2019-11, **Design of technical products and systems – Model of product design**.
- 
- [313] VDI 2221 Blatt 2:2019-11, **Design of technical products and systems – Configuration of individual product design processes**.
- 
- [314] DIN 820-2:2020-03, **Standardization – Part 2: Presentation of documents**.
- 
- [315] VDA, **Automotive VDA-Standardstruktur Komponentenlastenheft**. 1st ed., Berlin: VDA-QMC, 2007.
- 
- [316] Czarny, D. et al. (2020) **Project 2 Standards of the Future**. Pilot Petroleum sector. CCMC-Report (February).
- 
- [317] M. Schacht, Normen-Management, in: B. Bender, K. Gericke (Ed.) **Pahl/Beitz Konstruktionslehre**. 9th ed., Wiesbaden: Springer Vieweg, 2020.
- 
- [318] DIN SPEC 92001-2:2020-12, **Artificial Intelligence - Life Cycle Processes and Quality Requirements – Part 2: Robustness**.
-



9

List of authors

Dr. Rasmus Adler, Fraunhofer IESE

Thomas Andersen, Andersen Marketing KG

Marie Anton, Bundesverband der Arzneimittel-Hersteller e. V.

Dr. Andreas Aschenbrenner, Siemens AG

Yasmeen Babar, regio iT – Gesellschaft für Informations-  
technologie mbH

Adam Bahlke, Motor-AI

Dr. Markus Bautsch, Stiftung Warentest

Nikolas Becker, Gesellschaft für Informatik e. V. (GI)

Justus Benning, FIR e. V. an der RWTH Aachen

Bogdan Bereczki, ARGO AI

Bastian Bernhardt, IABG mbH

Dr. phil. Marija Bertovic, Bundesanstalt für Materialforschung  
und -prüfung (BAM)

Katharina Berwing, FIR e. V. an der RWTH Aachen

Tarek R. Besold, PHD, neurocat GmbH

Thordis Bethlehem, Berufsverband Deutscher  
Psychologinnen und Psychologen e. V. (BDP)

Paul Beyer, FSD Fahrzeugsystemdaten GmbH

Jörg Bienert, Bundesverband Künstliche Intelligenz e. V.

Antonio Bikic, Ludwig-Maximilians-Universität

Dr. Alexander Bode, CONABO GmbH

Jürgen Bönninger, FSD Fahrzeugsystemdaten GmbH

Dr. Julia Borggräfe, Bundesministerium für Arbeit und  
Soziales (BMAS)

PD Dr. med. Ulrich Bork, Universitätsklinikum Carl Gustav  
Carus an der Technischen Universität Dresden

Matthias Brand, MBDA Deutschland GmbH

Michael Brolle, Rembe GmbH

Dr. Joachim Bühler, Verband der TÜV e. V.

Dr. Aljoscha Burchardt, Deutsches Forschungszentrum für  
Künstliche Intelligenz GmbH (DFKI)

Prof. Dr. Armin Cremers, Rheinische Friedrich-Wilhelms-  
Universität Bonn

Stephanie Dachsberger, Plattform Lernende Systeme

Sharam Dadashnia, Scheer PAS Deutschland GmbH

Prof. Dr. Markus Dahm, IBM

Susanne Dehmel, Bitkom e. V.

Dr. Peter Deussen, Microsoft Deutschland GmbH

Verena Dietrich, imbus AG

Juergen Diller, Huawei

Alexander Dobert, Datenschutz Dobert

Jannis Dörhöfer, Verband der TÜV e. V. (VdTÜV)

Rebecca Ebner, acatech – Deutsche Akademie der  
Technikwissenschaften

Ralf Egner, Deutsche Akkreditierungsstelle GmbH (DAkkS)

Matthis Eicher, TÜV Süd

Patrik Eisenhauer, Collaborating Centre on Sustainable  
Consumption and Production gGmbH (CSCP)

Kentaro Ellert, PricewaterhouseCoopers GmbH

Filiz Elmas, DIN e. V.

Dr. rer. nat. Stefan Elmer, Festo SE & Co. KG

Jacques Engländer, FIR e. V. an der RWTH Aachen



Dr. Matthias Fabian, Landesärztekammer Baden-Württemberg

Patrik Feth, SICK AG

Marc Fliehe, Verband der TÜV e.V. (VdTÜV)

Prof. Dr. Alexis Fritz, Katholische Universität

Dr. Martina Frost, ifaa – Institut für angewandte Arbeitswissenschaft e.V.

Andreas Fuchsberger, Microsoft Deutschland GmbH

Philip Gallandi, NWB e.V.

Michael Gamer, Technische Universität Kaiserslautern

Prof. Dr. Dagmar Gesmann-Nuissl, Technische Universität Chemnitz

Wolfgang Gies, DVGW Deutscher Verein des Gas- und Wasserfaches e.V.

Marius Goebel, Spherity GmbH

Jan Götze, Airbus

Stephan Griebel, Siemens Mobility GmbH

Yvonne Gruchmann, Wirtschaftsförderung Land Brandenburg GmbH

Norman Günther, Technische Hochschule Wildau

Viktoria Hasse, Bundesverband Gesundheits-IT – bvitg e.V.

Marc Hauer, Technische Universität Kaiserslautern

Dr. Dirk Hecker, Allianz Big Data und Künstliche Intelligenz

Prof. Roland Heger, PhD, Integrata-Stiftung für humane Nutzung der Informationstechnologie

Jürgen Heiles, Siemens AG

Prof. Dr.-Ing. Michael Herdy, inpro Innovationsgesellschaft für fortgeschrittene Produktionssysteme in der Fahrzeugindustrie

Thorsten Hermann, Microsoft Deutschland GmbH

Dr. Sven Herpig, Stiftung Neue Verantwortung e.V.

Dr. Stefan Heumann, Stiftung Neue Verantwortung e.V.

Dr. Wolfgang Hildesheim, IBM Deutschland GmbH

Dr. Lukas Höhndorf, IABG mbH

Maximilian Hösl, Lernende Systeme – Die Plattform für Künstliche Intelligenz

Taras Holoyad, Bundesnetzagentur

Dr. Kristian Höpping, FSD Fahrzeugsystemdaten GmbH

Stephan Höppner, Atos Information Technology

Oliver Jähn, Landesärztekammer Brandenburg

Prof. Dr.-Ing. habil. Thomas Jürgensohn, HFC Human-Factors-Consult GmbH

Johannes Kahlhoff, Phoenix Contact GmbH Co. KG

Klaus Kaufmann, GS1 Germany GmbH

Michael Kayser, idox compliance

Dr. Till Klein, Initiative for applied artificial intelligence

Klaus Kleine Büning, TÜV Nord InfraChem GmbH

Marco Knödler, YNCORIS GmbH & Co. KG

Prof. Dr. habil. Jana Koehler, Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI)

Dr. Sergii Kolomiichuk, Fraunhofer-Institut für Fabrikbetrieb und -automatisierung IFF

Roman Konertz, FernUniversität in Hagen

Roland Kossow, CyberTribe®

Tobias Krafft, Technische Universität Kaiserslautern

Dirk Kretzschmar, TÜViT GmbH

Dr. Anna Kruspe, Deutsches Zentrum für Luft- und Raumfahrt (DLR)

Michael Krystek, Physikalisch-technische Bundesanstalt

Mark Küller, Verband der TÜV e. V. (VdTÜV)

Matthias Kuom, Europäische Kommission

Dr. Jens Lachenmaier, Universität Stuttgart

Holger Laible, Siemens AG

Philipp Lämmel, Fraunhofer-Institut für Offene Kommunikationssysteme FOKUS

Fredi Lang, Berufsverband Deutscher Psychologinnen und Psychologen e. V.

Dr.-Ing. Christoph Legat, HEKUMA GmbH

Dr. Olga Levina, FZI Forschungszentrum Informatik

Matthias Lieske, Hitachi Europe GmbH

Georg Ludwig Lindinger, Universität Bayreuth

Daniel Loevenich, Bundesamt für Sicherheit in der Informationstechnik (BSI)

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)

Christian Märkel, WIK GmbH

Prof. Dr. Klaus Mainzer, TUM Senior Excellence Faculty, Technische Universität München

Angelina Marko, Fraunhofer-Institut für Produktionsanlagen und Konstruktionstechnik IPK

Dr. Erik Marquardt, Verein Deutscher Ingenieure e. V. (VDI)

Jan de Meer, Hochschule für Technik und Wirtschaft Berlin (HTW Berlin)

Carsten Mehrrens, Volkswagen AG

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

Martin Meyer, Siemens Healthcare GmbH

Dr. Michael Mock, Fraunhofer-Institut für Intelligente Analyse und Informationssysteme IAIS

Prof. Dr. Andreas Mockenhaupt, Hochschule Albstadt-Sigmaringen (Albstadt-Sigmaringen University)

Thomas Möller, Bundesverband Gesundheits-IT – bvitg e. V.

Michael Mörike, Integrata-Stiftung für humane Nutzung der Informationstechnologie

Edeltraud Mörl, Dachverband für Technologen/-innen und Analytiker/-innen in der Medizin Deutschland e. V.

Andreas Müller, Schaeffler Technologies AG & Co. KG

Tobias Nagel, Fraunhofer-Institut für Produktionstechnik und Automatisierung IPA

Jan Noelle, Rettungsdienst-Kooperation in Schleswig-Holstein gGmbH

Dr. Shane O'Sullivan, Universidade de São Paulo, Brazil

Michael Paul, Safran S.A.

Fabian Petsch, Bundesamt für Sicherheit in der Informationstechnik (BSI)

Dr. Christoph Peylo, Robert Bosch GmbH

Christoph Pogorelow, IBM Deutschland GmbH

Frank Poignée, infoteam Software AG

Dr. Maximilian Poretschkin, Fraunhofer-Institut für Intelligente Analyse und Informationssysteme IAIS

Alexander Rabe, eco – Verband der Internetwirtschaft e. V.

Golo Rademacher, Bundesministerium für Arbeit und Soziales (BMAS)	Dr. Kinga Schumacher, Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI)
Dr. Frank Raudszus, Bundesnetzagentur	Silvio Schwarzkopf, FSD Fahrzeugsystemdaten GmbH
Ludwig von Reiche, NVIDIA ARC GmbH	Peter Seeberg, asimovero.ai
Janis Reinelt, AICURA Medical GmbH	Jan Seitz, Technische Hochschule Wildau
Axel Rennoch, Fraunhofer-Institut für Offene Kommunikationssysteme FOKUS	Aydin Enes Seydanlioglu, Robert BOSCH GmbH
Dr. Mathias Riechert, BMW Group	Annegrit Seyerlein-Klug, secunet Security Networks AG
Dr. Patrick Riordan, Siemens AG	Dr. Reiner Spallek, IABG mbH
Markus Röhler, Fraunhofer-Institut für Gießerei-, Composite- und Verarbeitungstechnik IGCV	Dr. Thomas Stauner, BMW Group
Dr. Miriam Ruf, Fraunhofer-Institut für Optronik, Systemtechnik und Bildauswertung IOSB	Andreas Steier, Deutscher Bundestag
Dr. Gerhard Runze, imbus AG	Dr. Reinhard Stolle, Argo AI
Ingo Sawilla, TRUMPF Werkzeugmaschinen GmbH + Co. KG	Dr. Manfred Stoyke, Bundesamt für Verbraucherschutz und Lebensmittelsicherheit (BVL)
Dr. med. Henning Schaefer, Ärztekammer Berlin	Prof. Dr. Karolina Suchowolec, Technische Hochschule Köln
Dr. Stefan Schaffer, Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI)	Julia Szelag, AWW Arbeitsgemeinschaft für wirtschaftliche Verwaltung e. V.
Elias Schede, PricewaterhouseCoopers GmbH (PwC)	Steffen Tauber, evia
Christopher Scheel, SCHUFA Holding AG	Dr. Nikolay Tcholtchev, Fraunhofer-Institut für Offene Kommunikationssysteme FOKUS
Prof. Dr.-Ing. Ina Schieferdecker, Bundesministerium für Bildung und Forschung (BMBF)	Martin Tettke, Berlin Cert GmbH
Raoul Schönhof, Fraunhofer-Institut für Produktionstechnik und Automatisierung IPA	Dr. Volker Treier, Deutscher Industrie- und Handelskammertag e. V.
Dr. Thomas Schmid, Universität Leipzig	Dr. Denise Vandeweyer, UnternehmerTUM GmbH
Dr. Jörg Schneider, Bundesnetzagentur	Dr. Silvia Vock, Bundesanstalt für Arbeitsschutz und Arbeitsmedizin (BAuA)
Martin A. Schneider, Fraunhofer-Institut für Offene Kommunikationssysteme FOKUS	Dr. Thomas Vögele, Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI)
MinDirig Stefan Schnorr, BMWi	Roland Vogt, Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI)

Kirsten Wagner, DVGW Deutscher Verein des Gas- und Wasserfaches e. V.

Prof. Dr. rer. nat. Dr. h.c. mult. Wolfgang Wahlster, Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI)

Dr. Thomas Waschulzik, Siemens Mobility GmbH

Prof. Dr.-Ing. Dieter Wegener, Siemens AG

Prof. Dr. Johann Wilhelm Weidringer, Bundesärztekammer (Bayer. Landesärztekammer)

Wei Wei, IBM Deutschland GmbH

Dr. Frank Werner, Software AG

Martin Westhoven, Bundesanstalt für Arbeitsschutz und Arbeitsmedizin (BAuA)

Dr. Johannes Winter, Plattform Lernende Systeme

Christoph Winterhalter, DIN e. V.

Raoul Wintjes, DSLV Bundesverband Spedition und Logistik e. V. (DSLVL)

René Wöstmann, RIF e. V. Institut für Forschung und Transfer

Yuanyuan Xiao, BTC AG

Jens Ziehn, Fraunhofer-Institut für Optronik, Systemtechnik und Bildauswertung IOSB

**10**

**Further working group  
members**

Fabian Anzmann, HKI Industrieverband Haus-, Heiz- und Küchentechnik e.V.

Eberhard Becker, DEMAG CRANES & COMPONENTS GMBH

Dr. Andreas Binder, Samson AG

Miika Blinn, Verbraucherzentrale

Thomas Boué, BSA | The Software Alliance

Gebhard Bouwer, TÜV Rheinland Industrie Service GmbH

Dr. Konstantin Böttinger, Fraunhofer-Institut für Angewandte und Integrierte Sicherheit AISEC

Dr. Alfonso Caiazzo, WIAS Berlin

Beatriz Cassoli, Technische Universität Darmstadt

Klaus Däßler, Gesellschaft für Mathematische Intelligenz

Thierry Declerck, Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI)

Alberto Diaz-Durana, HEDERA Sustainable Solutions GmbH

Dr. Markus Dicks, Bundesministerium für Arbeit und Soziales (BMAS)

Jörg Dubbert, VDI/VDE Innovation + Technik

Heiko Ehrich, TÜV NORD Mobilität GmbH & Co. KG IFM

Bernd Eisemann, Munich RE

Karl-Ludwig Elfira Blumenthal, Siemens AG

Filiz Elmas, DIN e.V.

Alexandra Engelt, DIN e.V.

Prof. Dr. Stefan Evert, Friedrich-Alexander-Universität Erlangen-Nürnberg

Dr. med. Matthias Fabian, Landesärztekammer Baden-Württemberg

Dr.-Ing. Stephan Finke, DAkkS – Deutsche Akkreditierungsstelle

Robert Freund, Bundesanstalt für Materialforschung und -prüfung (BAM)

Egbert Fritzsche, VDA

Enno Garrelts, Universität Stuttgart

Dr. Moritz Gerlach, Bundesministerium für Bildung und Forschung (BMBF)

Marian Gläser, brighterAI

Richard Goebelt, Verband der TÜV e.V.

Benedikt Grosch, Technische Universität Darmstadt

Dr. Oliver Grün, BITMi

Dr. Thilo Hagendorff, Universität Tübingen

Andreas Hartl, Bundesministerium für Wirtschaft und Energie (BMWi)

Manfred Hefft, Domino Deutschland GmbH

Stefan Herr, innogy SE

Dr. Klaus Hesselmann, yourexpertcluster

Max Hofmann, Volkswagen AG

Dr. rer. pol. Reiner Hofmann, Universität Bayreuth

Dr.-Ing. Gerhard Imgrund, VDE

Dr. Andreas Jedlitschka, Fraunhofer-Institut für Experimentelles Software Engineering IESE

Stephan Jenzen, Airbus

Prof. Dr. Christian Johner, Johner Institut GmbH

Szilvia Kalmar, Erste Lesung GmbH

Ninmar Lahdo, VDE

Johannes Koch, VDE

Stephan Krähnert, VDA

Danny Lubosch, Gefertec GmbH

Prof. Dr. Christoph Lütge, Technische Universität München

Dr. Oliver Maguhn, Munich RE

Johannes Melzer, Deutscher Industrie- und Handelskammertag e. V.

Dirk Michelsen, IBM Deutschland GmbH

Sebastian Micus, Deutsches Institut für Textil- und Faserforschung Denkendorf (DITF)

Stefan Mitterer, OELCHECK GmbH

Andreas Möller, ADVES GmbH & Co. KG

Prof. Dr. Jürgen Mottok, Ostbayerische Technische Hochschule Regensburg

Gert Nahler, Samson AG

Dr. med. Felix Nensa, Universitätsklinikum Essen

Thomas Niessen, Kompetenznetzwerk Trusted Cloud e. V.

Luis Oala, Fraunhofer-Institut für Nachrichtentechnik HHI

Sebastian von Oppen, Architektenkammer Berlin

Dr.-Ing. Christian Peter, BioArtProducts GmbH

Dr. Georg Plasberg, SICK AG

Christoph Preuße, Berufsgenossenschaft Holz und Metall

Dr. Gerald Quitterer, Bayerische Landesärztekammer

Dr. Martin Radtke, Bundesanstalt für Materialforschung und -prüfung (BAM)

Dr. phil. Georg Rehm, Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI)

Prof. Dr.-Ing. Jörg Reiff-Stephan, Technische Hochschule Wildau

Guido Reusch, eapr GmbH

Christina Rode-Schubert, trend2ability

Laurent Romary, INRIA

Jan Rösler, DIN e. V.

Nils Röttger, imbus AG

Dennis Scheuer, IBM Deutschland

Robin Schlenga, Ramboll Management Consulting GmbH

Prof. Dr. Ralf Schnieders, Hochschule für Technik und Wirtschaft Berlin

Andreas Schumann, Bundesverband der Kurier-Express-Post-Dienste

Dr.-Ing. Dennis Schütte, Still GmbH

Philip Sperl, Fraunhofer-Institut für Angewandte und Integrierte Sicherheit AISEC

Johann-Christoph Stang, Bundesanstalt für Materialforschung und -prüfung (BAM)

Patrick Stanula, Technische Universität Darmstadt

Timo Strohmann, Technische Universität Braunschweig

Klaus Strumberger, Morgen & Morgen GmbH

Denis Suarsana, Bundesvereinigung der Deutschen Arbeitgeberverbände (BDA)

Claudia Tautorius, Verband der TÜV e. V.

Dr.-Ing. Max Ungerer, PROSTEP AG

Dr. David Urmann, VDE

Dirk Walther, Fahrzeugsystemdaten GmbH

Dr. Markus Wenzel, Fraunhofer-Institut für Nachrichtentechnik HHI

Dr. Jing Xiao, Continental Automotive

Salah Zayakh, REWE digital GmbH

Dr. Carlos Zednik, Universität Magdeburg

Dr. Thomas Zielke, Bundesministerium für Wirtschaft und Energie (BMWi)



**11**

**Annex**

## 11.1 Glossary

Areas:

- 1 Artificial Intelligence (only general terms)
- 2 Characteristics of AI systems
- 3 Characteristics of data
- 4 Methods and techniques
- 5 Machine learning (also neural networks)

**Table 14:** Glossary

Areas	German	Alternative German	English	Description and source <sup>30</sup>
1	Agent	Softbot	agent	An agent is generally defined as a software or hardware unit that processes information and produces an output from an input.
3	Aktualität		currency	The degree to which data has attributes that are of the right age in a specific context of use. [88]
5	Angeleitetes maschinelles Lernen	überwachtes Lernen	supervised machine learning	Machine learning technique based on the use of pre-classified data <sup>31</sup>
1	Automatisierung		automation	Replacement of manual activities by computerized methods [294]
2	Autonomie		autonomy	Ability of a system to perform tasks based on its internal state and environment without human intervention <sup>32</sup> [295]
1	Big Data		Big Data	Data that is too extensive, too complex, too short-lived or too weakly structured to be evaluated using conventional methods of data processing. <sup>33</sup>
3	Datenqualität		data quality	Quality related to data [88] <sup>34</sup> Degree to which the characteristics of data satisfy stated and implied needs when used under specified conditions [88]
3	Datenschutz		data protection	Privacy/data protection (data security) refers to the collection and processing of personal data according to the relevant regulations such as the EU General Data Protection Regulation. For example, persons in Europe have the right to have their private data adequately protected against IT attacks.

<sup>30</sup> Where no source is given, the description was written by the team of authors of this Roadmap. If the source is an unpublished working paper, it is indicated in a footnote, otherwise by reference.

<sup>31</sup> Source: ISO/IEC CD 22989 (open project, not published).

<sup>32</sup> In the ISO description the term “automatic” but not “autonomy” (= “in one’s own name” and therefore sanctionable and under one’s own responsibility) is described. A counterexample is a wildlife camera with fully automatic setting and recording, which is certainly not considered autonomous. Incidentally, an autonomous system is probably more than just a cognitive system.

<sup>33</sup> Source: ISO/IEC WD 20546 (Draft as of 2019-08-04, open project, not published.) ISO note to entry: Big data is commonly used in many different ways, for example as the name of the scalable technology used to handle big data extensive data sets.

<sup>34</sup> Derived from DIN EN ISO 9000:2015 [105]. Data quality is described in characteristics or dimensions, for example with inherent characteristics such as accuracy and completeness, or system-dependent characteristics such as availability and recoverability.

Areas	German	Alternative German	English	Description and source <sup>30</sup>
5	Deep Neural Network		Deep Neural Network	Neural network that has further, hidden node layers in addition to the input and output layer (cf. Deep Learning)
2	Erklärbarkeit	Nachvollziehbarkeit	explainability	Property of an AI system that factors that have led to an automated decision of the system can be understood by a human <sup>35</sup>
4	Experten-system	expert system		Often rule-based system based on symbolic knowledge processing. Example: if-then rules. → Symbolic, formal representation of knowledge in AI systems. Conclusion, using logic to derive new knowledge from formal knowledge
3	Genauigkeit		accuracy	The degree to which data has attributes that correctly represent the true value of the intended attributes of a concept or event in a specific context of use. [88]
2	Grad der Zuverlässigkeit		dependability	Ability to execute in the required manner and at the required time [105]
4	Inferenz	logisches Schließen	inference reasoning	Rule-based reasoning; often used in expert systems
1	KI-Komponente		AI component	System component that uses Artificial Intelligence
1	KI-Modul		AI module	Software module that implements algorithms [86]
1	KI-System		AI system	System that uses artificial Intelligence <sup>36</sup>
1	Kognitives System		cognitive system	Adaptable system with interfaces to the digital world and the environment, which can perceive things automatically, relate to and understand contexts and draw conclusions and learn from them in order to solve and master tasks.
5	Kontinuierliches Lernen		continuous learning	Incremental training of an AI system, which takes place continuously in the production environment of the system <sup>37</sup>
2	Kontrollierbarkeit	Steuerbarkeit	controllability	Ability of a human operator to intervene in a timely manner in the functioning of a system <sup>38</sup>
5	Lerndaten	Trainingsdaten	training data	Data used to train a model <sup>39</sup>
1	Lernendes System		learning system	Learning systems are machines, robots and software systems that independently perform abstractly described tasks on the basis of data that serve as a basis for learning, without each step being specifically programmed by humans. To solve their task, they use models trained by learning algorithms. With the help of the learning algorithm, many systems can continue learning during operation: They improve the pre-trained models and expand their knowledge base. [296]

35 Source: ISO/IEC CD 22989 (open project, not published)

36 Source: *ibid.*

37 Source: *ibid.*

38 Source: *ibid.*

39 Source: *ibid.*

Areas	German	Alternative German	English	Description and source <sup>30</sup>
4	Maschinelle Übersetzung		machine translation	Automatic translation of spoken or written natural language into another language by an AI system <sup>39</sup>
5	Maschinelles Lernen		machine learning	Technology that enables a system to learn from data and interactions
1	Modell		model	physical, mathematical, or otherwise logical representation of a system, entity, phenomenon, or process” [297]
5	Neuronales Netz	künstliches neuronales Netz, KNN	artificial neural net	Calculation network consisting of simple calculation elements and weighted relations between these elements, whose input-output function is determined by the interaction of the network elements <sup>40</sup>
1	Roboter		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 the environment itself. <sup>42</sup>
4	Robotik		robotics	Discipline that deals with the construction of robots.
2	Robustheit		robustness	Ability of a system to fulfil its function under any circumstances. <sup>43</sup>
2	Safety	Sicherheit	safety	Refers to the expectation that under certain circumstances a system will not lead to a state in which human life, health, property or the environment are endangered.
5	Schwach überwachttes maschinelles Lernen	teilüberwachtes Lernen	semi-supervised machine learning	Degree of temporal validity of data that is relevant in a certain application context. <sup>44</sup>
1	Schwache KI		narrow (or weak) AI	AI system designed for a specific purpose
2	Security	Sicherheit	security	Aims to prevent negative effects that a human or other machine can have on the AI module. Confidentiality, integrity and availability are the most important security objectives.
4	Semantische Berechnung	semantische Technologien	semantic computing	Technologies aimed at the representation and processing of knowledge <sup>45</sup>
1	Starke KI		general (or strong) AI	(Theoretical construct:) general intelligence that can set goals for itself <sup>46</sup>

40 Source: ibid.

41 Source: ibid.

42 Source: ibid.

43 Source: ISO/IEC WD 24029-1 (Draft as of 2019-10-30, open project, not published.)

44 Source: ISO/IEC WD 20546 (Draft as of 2019-08-04, open project, not published.) ISO note to entry: The training data for a semi-supervised learning task can include a majority of unlabelled inputs.

45 Source: ISO/IEC CD 22989 (open project, not published).

46 s. also [43]

Areas	German	Alternative German	English	Description and source <sup>30</sup>
5	Tiefes Lernen	mehrschichtiges Lernen	deep learning	Machine learning technique based on artificial neural networks with several hidden layers
5	Trainiertes Modell	angelerntes Modell	trained model	Model resulting from machine learning <sup>47</sup>
5	Training		training	Process for establishing or improving models using machine learning <sup>48</sup>
2	Transparenz		transparency	Open, complete, understandable and accessible presentation of information on functional aspects of an AI system. This includes, among other things, the explainability of the AI system (e.g. neural networks), the traceability of the data protection concept and information on quality assurance processes during development.
5	Unüberwachtes maschinelles Lernen	unbeaufsichtigtes maschinelles Lernen	unsupervised learning	Machine learning technique based on the use of non-classified data <sup>49</sup>
2	Verständlichkeit		understandability	
5	Verstärkendes Lernen	bestärkendes Lernen	reinforcement learning	Technique of machine learning based on the positive or negative evaluation of attempts of a system <sup>50</sup>
2	Vollständigkeit		completeness	Degree to which data associated with an entity has values for all attributes of this entity and for all related entities. [88]
4	Wissensrepräsentation		knowledge representation	Representation of knowledge that can be used for an AI system, e.g. an expert system
3	Zugänglichkeit	Verfügbarkeit	accessibility, availability	

47 Source: ISO/IEC CD 22989 (open project, not published.)

48 Source: *ibid.*

49 Source: *ibid.*

50 Source: *ibid.*

## 11.2 Philosophical foundations of ethics

In order to be able to deal with ethics in relation to AI systems, one should deal with the basics of philosophy and thus its special field of ethics in our cultural area. In general, philosophy (ancient Greek φιλοσοφία, latinized **philosophia**, literally “love of wisdom”) is the attempt to fathom, interpret and understand the world and human existence. Philosophy differs from other scientific disciplines in that it is often not limited to a specific field or methodology, but is characterized by the nature of its questions and its particular approach to its manifold subject areas. (see Wikipedia article). There is no universal philosophical method; on the contrary, there are a multitude of them, which in turn adhere to certain currents, such as hermeneutics, which is a generally accepted method in the humanities. Hermeneutics denotes something like an understanding interpretation of documents of consciousness, a method of interpretive art of interpretation, but also altogether a philosophical theory of understanding in its premises, foundations and results. Because of this breadth, the term hermeneutics can be found in a wide variety of theoretical contexts; conversely, critics of hermeneutics often do not know where to start. Dialectic is another method; a philosophical method that questions the position from which it originates through opposing assertions and seeks to gain a higher kind of knowledge by synthesizing both positions.

The philosophical approach was supplemented by Sigmund Freud (1856-1939), who placed a noteworthy importance for a change in the view of humans. Freud was an Austrian neurophysiologist, psychoanalyst, cultural theorist and critic of religion. He is considered one of the most influential thinkers and world-changers of the 20th century, especially due to his foundation of psychoanalysis. His theories and methods are still discussed, applied and criticized today. The reason why psychoanalysis was so ground-breaking was that it allowed, for the first time, an access to the unconscious and thus to the actions of people and the contemplation of being. Psychoanalysis later developed into the various schools of psychology.

The “Lexikon Philosophie” [298] provides a good initial definition of ethics: “it is defined as that branch of philosophy which deals with the preconditions and evaluation of human action and is the methodical reflection on morality. At the centre of ethics is specifically moral action, especially with regard to its justifiability and reflection (ethics describes and critically evaluates morality). (...) Ethics and its neighbouring disciplines (e.g. philosophy of law, philosophy of state and

social philosophy) are also summarized as ‘practical philosophy’, since they deal with human action”.

“In ethics, the sub-areas of normative ethics, metaethics and applied ethics can be distinguished. Normative ethics develops evaluative theories of desirable action. The subject of metaethics is normative ethics itself – it questions, for example, its basic assumptions or analyzes the processes of normative ethics. Applied ethics focuses on specific areas of life and tries to reflect and shape them in consideration of normative ethics and metaethics”. [299]

There are various modern ethical approaches that can be applied to artificial intelligence, as well as philosophers and works dealing with ethical AI. What is interesting here is that, for the first time, ethics refers to a machine rather than to humans alone.

Although an AI ethics is not yet clearly and conclusively defined, it can certainly be located in the field of applied ethics. When it comes to the ethical considerations concerning the technical aspects of AI, it has strong links to the cross-sectional area of machine ethics [300]; when it comes to the socio-technical and economic aspects, it has strong links to economic ethics. In addition, it will have regular references to the vertical areas of applied ethics, such as bio- or medical ethics, whenever area-specific ethical considerations are to be updated in light of the AI phenomenon.

In addition, “ethics” have evolved in many application areas as a result of the specific challenges of individual fields of application, such as the following:

- The ethics of law, which is a part of law as well as part of applied philosophy. It is distinguished in two basic aspects from other “ethics of science”. “On the one hand, ethical and moral norms do not meet here with a section of reality structured more strongly by facts or laws of nature – such as nature, technology and medicine, for example – but with the law, as an order of concepts and norms that fundamentally [...] normatively arch over and shape and conceptually structure reality.” On the other hand, legal ethics has been dealing with these questions ever since human societies and their philosophical considerations have existed [300].
- Medical ethics deals with the moral standards that should apply to the health care system. It has evolved from doctor’s ethics, but affects all persons, institutions and organizations working in the health care system and, last but not least, the patients. Closely related disciplines

are medical humanities and bioethics. The fundamental values are the well-being of the human being, the prohibition to harm (lat.: “nemini noceri!”) and the right to self-determination of patients (principle of informed consent), more generally the principle of human dignity [302].

At the present time, ethics in the sense of a general understanding are assumed to ensure that AI systems ultimately follow our legal rules in their application, as well as “responsibly” dealing with human values of our society. These can be approached also by means of ethical criteria of professional associations (e.g. High Level Group of the EU, or the Platform Learning Systems).

Especially in the press and media – and thus in the awareness of the public – as well as partly in research, ethical dilemmas are often raised in connection with the use and the rapid further development, as well as the already used but also targeted use of AI, e.g. the question of the “behaviour” of an automated vehicle in critical traffic situations. Should the AI system choose one person over a group in case of danger to life and limb?

When talking about ethics in relation to AI, addressing ethical/moral dilemmas is a necessity, even if the goal must be to prevent them in advance, i.e. to design the autonomous machine in a way that would not endanger anyone. Or one “feeds” the AI system in advance with the ethical values of our culture.

An **ethical dilemma** is a situation where a decision to act is required even though every possible option for action, including non-action, inevitably violates an ethical postulate. Since the use of AI systems always involves a certain degree of loss of control and thus creates risks, there is always an implicit trade-off between the potential dangers and the potential benefits of AI. This is particularly critical if the health or life of people can be potentially endangered. The German Federal Constitutional Court already ruled out the possibility of “offsetting” human lives in 2006 [303]. The Ethics Commission confirmed this in principle for automated vehicles in 2017, but opened it up as follows: “**General programming to reduce the number of personal injuries may be acceptable**”.

The **Principle of Double Effect** deals with the question of moral responsibility when a morally good decision has an (unintended) ethically bad side effect. A special case of the principle of double effect is **Dual Use**. This raises the question

of whether a developer or manufacturer is responsible for a harmful use unintended or prohibited by them. The term originates from the export control of products that can be used simultaneously for civil and military purposes, but is also applied to ethical dilemmas.

Both the excursus into the concrete ethical question of the dilemmas quasi at the end of the chain, and the consideration of the ethical-philosophical overall development of our society make clear that in the development and use of AI, there are no superior and universally applicable ethics from which one can derive valid rules. The above-mentioned philosophical and ethical development of our society, especially with regard to general values, makes it very clear that the European cultural area must develop and derive a suitable framework that is compatible with our laws from (Western) values and norms.

### 11.3 SafeTRANS Roadmap

The Roadmap of the SafeTRANS working group with the title “Safety, Security, and Certifiability of Future Man-Machine Systems” provides a model (see Figure 30) for the explainability of the complexity of human-machine systems and thus is a good example for the integration of safety and security. According to this, it should be possible to characterize AI systems interacting with people under the five aspects of “system strength”, “context”, “cooperation”, “responsibility & reflection” and “integrity & certification”. The aspects are assigned target vectors with scales for specific processes, methods and capabilities.

The thematic congruence to this Standardization Roadmap AI comes via the vector “responsibility and reflection” and “integrity and certification”. They show how decisions of the system are weighed up on a legal, ethical and moral level (“responsibility and reflection”) and how a decision can be assessed according to consistency, trustworthiness and risk classification (“integrity and certification”). The other target vectors are described in more detail below. “System strength” is represented by degrees of autonomy, intelligence and evolution. Under “Context” an analysis of the human and physical environment of the system under investigation is proposed. “Cooperation” considers a support by further systems or a human intervention [304].

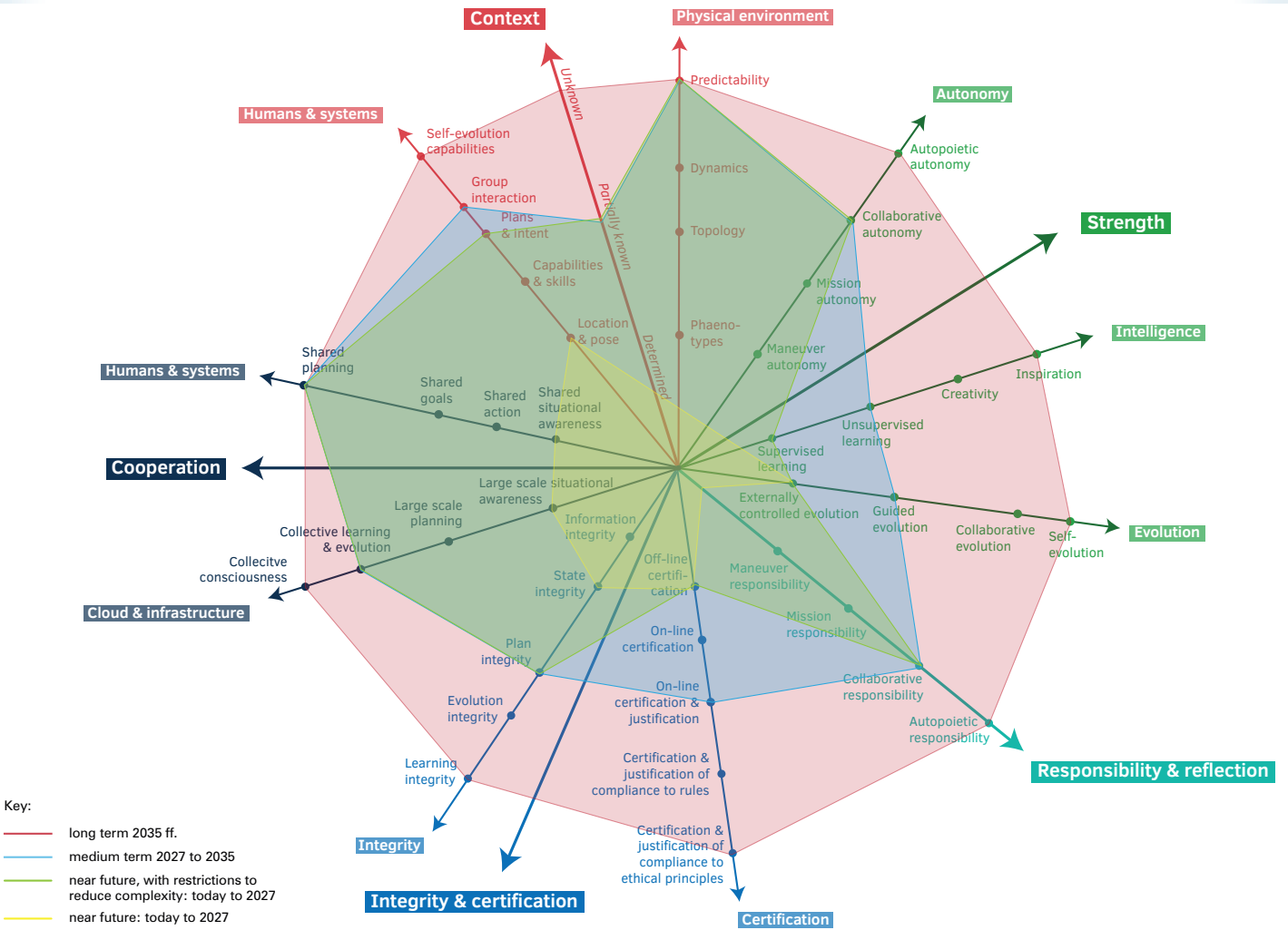


Figure 30: SafeTRANS Roadmap for the explainability of the complexity of human-machine systems [304]

### 11.4 SMART Standards – New design of standards for AI application processes

In this annex, the central question of what a new design of standards for AI user processes looks like and which technological approaches can be used in this context is examined in greater detail.

The procedures for providing granular standards information will vary (see also Chapter 5.2.5).

→ **Technology approach:** Existing standards documents are automatically indexed in **post-processing** without subject and number limitations and are automatically provided in granular “addressable” information units using semantic methods. The indexing accuracy is currently about 80 % compared to intellectually granularly prepared documents and thus meets the requirements of qualified users who can evaluate the disassembled

information offer professionally. But for downstream AI application processes this means that a validation of the accuracy of the partial information must be integrated. The drivers of this approach are “content management” and “content delivery”. Results in **Level 3** (with above mentioned limitations) **based on Level 2** are achievable.

→ **Bottom up approach:** When digitizing standards, a distinction can be made between a top-down and a bottom-up approach. Both approaches deal with questions of modularization, modelling and management of future standard content, but from different perspectives. Here, the top-down approach is characterized by the redesign of the actual standardization process and the question of how future digital standards must be structured, whereas the bottom-up approach deals with the transfer of already existing standard contents (“restructuring”) into a machine-executable knowledge representation. The development of smart standards requires both a top-down and a bottom-up approach. The drivers of the bottom-up



approach are “content management and delivery” and “content usage”. **Level 3 and Level 4** results are achievable for defined, delimited areas of application.

→ **Top down approach:** There can only be one reference document or “reference content” of the standard and this is the content that has been checked and approved by the responsible standards body, the “primary content”. As a rule, laws or contracts refer only to these and only this primary content is relevant in serious cases. So that the machine-readable standard content can also be primary content, the acquisition of the human-generated and -readable linguistic standard content must be carried out in preprocessing (in the sense of the standards development process) on the basis of a structure that allows the language, including the semantics it contains, to be unambiguously transformed into a machine-readable data structure (e.g. ontology) and vice versa. The drivers of this approach are “content creation” and “content usage”. **Level 4 results** are achievable.

**Processing sequence**

The different approaches can and should be pursued in parallel. The technology approach provides faster insights that can be used in the other approaches. In addition, the first – economically viable – customer solutions or prototypes and demonstrators are quickly developed, so that practical experience can be fed back. The bottom-up approach cannot

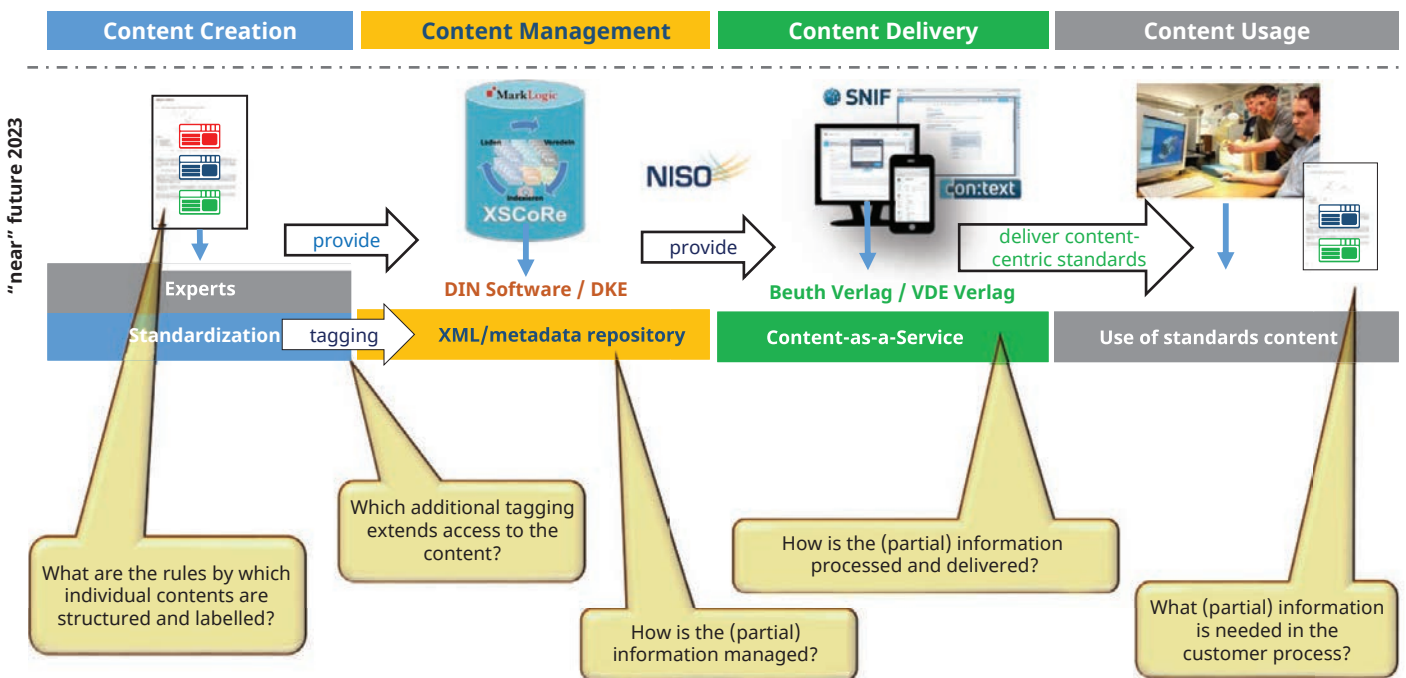
be suitable for structuring the very large and constantly growing worldwide stock of standards to the highest quality standards. But in this procedure, too, the idea is to start purposefully in order to gain experience. However, the post-processing of standards can be economical for concrete fields of application. The “top class” for the announced goal of achieving SMART standards with the highest quality requirements for AI application processes can only be the pursuit and implementation of a top-down method (preprocessing). This will require a great amount of effort.

**11.4.1 Use of granular content by means of the technology approach**

**Level 2 and 3**

The process for Levels 2 and 3 is – as for Level 1 – characterized by delimited traditional areas of responsibility. While this simplifies the implementation of solutions from an organizational point of view, it prevents the integrated overarching action that becomes mandatory for SMART standards at Level 4. The focus on IT-supported processes and their further development in “content management” and “content delivery” offers the chance to quickly arrive at concrete solutions that provide valuable input for Level 4. A basis for the necessary IT infrastructure is also laid. The main questions to be answered are listed in Figure 31.

**Level 2 and 3**



**Figure 31:** Level 2 and 3 process and relevant issues

New digital solutions for the application of standards based on XML technology have emerged or are currently being developed [305], [306]. Further solutions, which are possible due to structured content, will be developed soon. The status of the further developments is briefly described using examples.

For downstream AI application processes this means: A validation of the accuracy of the automatically determined (partial) information must be performed. Knowledge gained from experience can essentially support the evaluation.

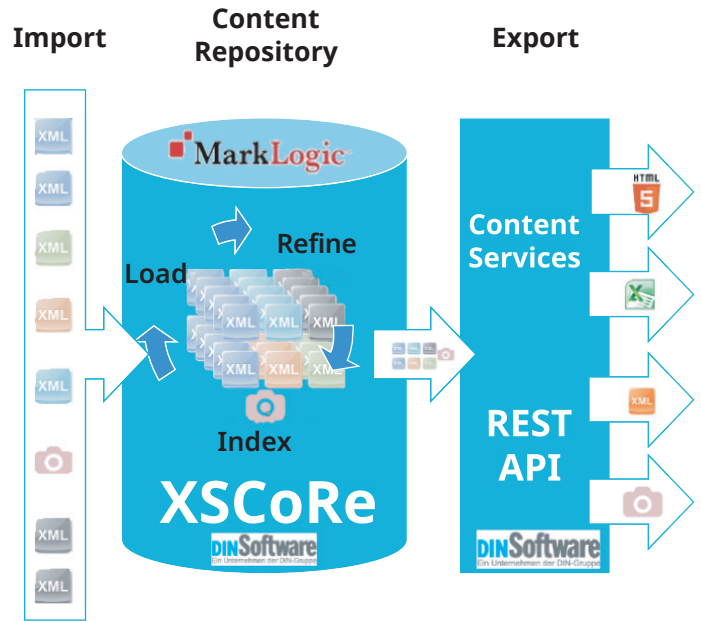
General rules for the description of (partial) information in standards and the methodological development of the exact places of use (sites of action, see Annex 11.4.3) are not yet available for this approach and must be developed. In order to provide AI application processes with (partial) information in a scalable way, appropriate specifications must be agreed upon.

**EXAMPLES OF DEVELOPMENTS IN “CONTENT MANAGEMENT” AND “CONTENT DELIVERY”**

**Example A: Key technology XML database and interfaces**

In 2016, DIN Software GmbH began converting the DIN Standards into XML (see Figure 32). At the same time an XML database was created [306]. The goals were and are the:

- Development of an “XML Semantic Content Repository” (XSCoRe) as a digital content search and delivery platform
- Central provision and management of metadata and standards content for DIN Group digital information and knowledge products
- Provision of interfaces to support the content management processes of the XML workflows of DIN and Beuth in order to connect them effectively to the XML repository
- Provision of a basis for development of “granular content-as-a-service based platform services” and thus acceleration of the digital transformation of the DIN Group’s business processes



**Figure 32:** Conversion of the DIN Standards collection into XML

XSCoRe is to be further developed to the extent that, for example, requirements can be mapped in RegIF format.

**Example B: Key technology XML application rules – NISO Z39.102-2017 [307]**

To ensure that standards have an identical XML format, a “Standard Tag Suite” was developed [308]:

- On 2017-10-09 recognized by ANSI as a US standard
- Ca. 30 standards publishers worked on it, including BSI, SFS, DIN, CEN, ISO, IEC, IEEE, ASTM, ASME.
- The defined “Tag Set” is the basis for the exchange and provision of XML standards (ISO, IEC, DIN, ASME, ...)
- DIN products use content in the NISOSTS format

At the same time it was agreed to further develop a “Standards-Specific Ontology Standard (SSOS)” in a NISO Working Group in 2020, see NISO Information [308].

The members of the National Information Standards Organization (NISO) have approved a new project to create a standards-specific ontology standard (short title: NISO SSOS). A working group will be formed to develop and standardize a high-level ontology to describe a limited number of core

concepts and relationships, initially focusing on the life cycle of standards.

This will facilitate the use of standards, support more consistent discovery and navigation within standards, and provide a foundation for other semantic applications, such as linked data, in the standards ecosystem. The agreement on an ontology enables standards publishers and distributors to continue to use existing investments in XML. It builds on existing work such as the NISO Standard Tag Suite, an ANSI/NISO standard that is a set of XML elements that provides a common format for the presentation and exchange of standard content, regardless of how the content is ultimately delivered to customers.

The Standards-Specific Ontology Standard (SSOS) provides the foundation required to move forward in key areas such as improved machine readability (Level 2). The associated content-related enhancement of the documents will also be evaluable, especially for AI-based applications. Since the project is currently still in a start-up phase, AI-related requirements should be named and – if possible – included.

**Example C: Key technology “Semantic standards information framework (SNIF)”**

With the “Semantic Standards Information Framework (SNIF)” a semantic indexing of standards has been realized [309]. In SNIF, the metadata from the DITR database and the standard texts are semantically indexed, thus achieving high quality results (see Figure 33).

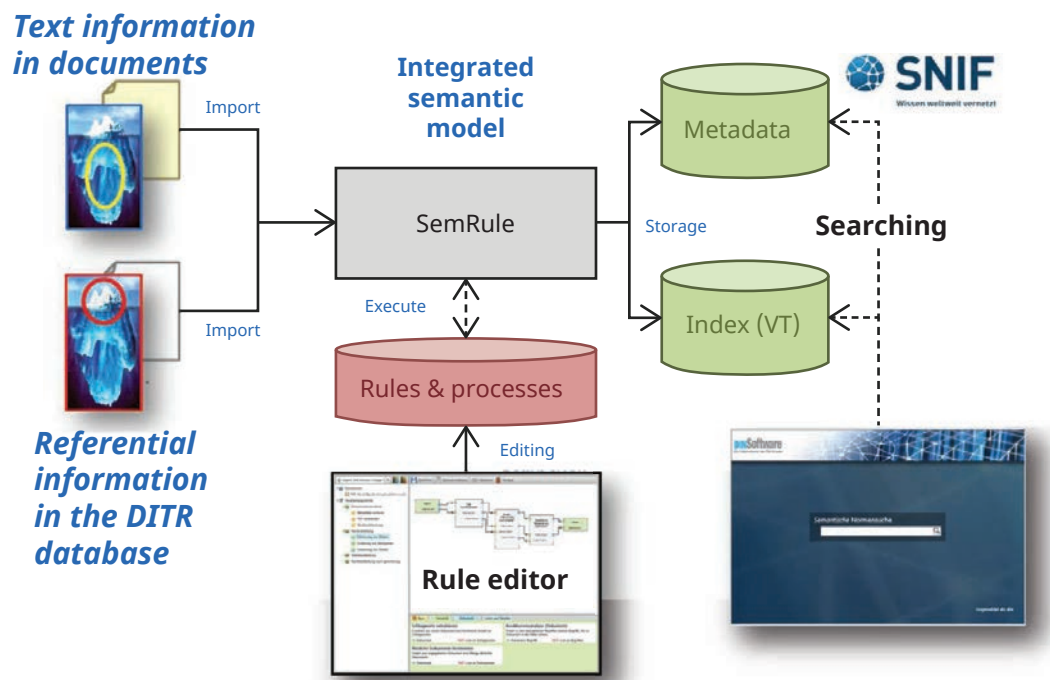
The user-friendly interface for data extraction and updating the database allows even employees without programming skills to set up and adapt processes for extracting relevant information from DIN standards as required using a rule editor.

Today,

- similarities between standards are systematically analyzed,
- various indexing and content enrichment services are provided, and
- a tailor-made meta-data service is made possible by SNIF.

SNIF is a basic technology whose extraction possibilities have not yet been fully exploited. In the AI project it is to be checked in which way the framework can be used.

Figure 33: SNIF basic technology



**Example D: Key technology “con:text”, to find distributed information and to show connections**

The service “con:text” was developed based on XML-converted documents and in compliance with the NISO STS, which can be linked to various standards management systems. The set of functions aims at a deeper understanding of the content, playing out correlations simultaneously and making them visible in a user-friendly way via numerous functions. In a further expansion stage, the creation of company-specific documents (e.g. company standards, technical delivery specifications) is supported in bidirectional interaction with DIN standards, see [Figure 34](#).

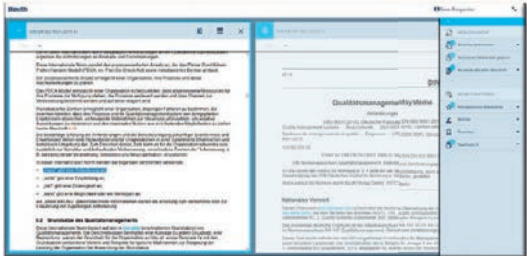
Data basis for con:text:

- Conversion of 35,000 DIN EN ISO documents into XML
  - ca. 4.5 million active links

The set of con:text functions reflects the needs of the user. Thus, an application know-how is created here that can be relevant for the functional formation of AI application processes. At the same time, the con:text application can benefit from the results of the AI project. Participation in the AI project should be made possible.

**Figure 34:** Functions in con:text

**con:text – Expanded functions in applications with the focus on “content”**



Indicating links

Can be integrated in collection

Online editing

Evaluation of interdependencies

Track changes

Redlines on the fly

Collaboration

Supervision at the granular level

Edit formulae in the editor

Indication of terms and requirements

Edit and insert tables

Access to “granular” elements (text, figures, tables, formulae)

**Example E: Key technology “extraction of standards content”**

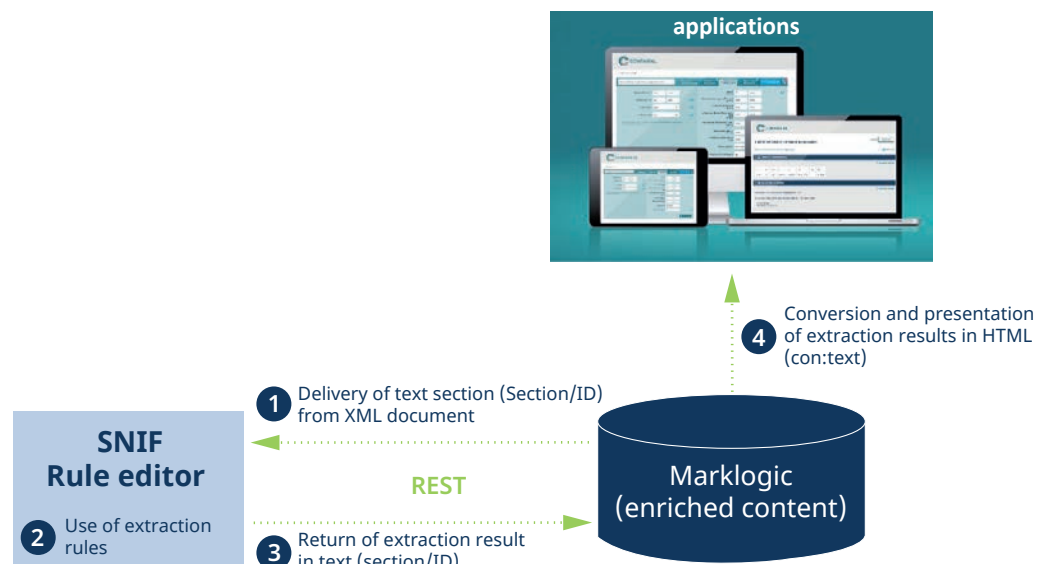
Based on XML-converted documents and the “Semantic Standards Information Framework (SNIF)”, a research method is being developed that can extract text passages from standards in a rule-based manner (see Figure 35).

- The content is retrieved from the XML Marklogic database via interfaces that are provided (usually individual clauses).
- The Semantic Standards Information Framework (SNIF) can receive this content in XML format via its own REST interfaces.

- The analysis and processing (recognition of requirements and recommendations) take place within the SNIF system.
- Marklogic receives via a REST interface the generated added value and delivers it to applications such as con:text.

The rules for information extraction are formulated according to customer requirements. Thus, an application know-how is created here that can be relevant for the functional formation of AI application processes. Participation in the AI project should be made possible.

**Figure 35:** Key technology “extraction of standards content”



**11.4.2 Bottom-up method – Post-processing of standards**

Five steps are necessary for implementing the bottom up approach – “Extraction”, “Modelling”, “Fusion and storage”, “Provision” and “Application” (see Figure 36) [310]. The question arises as to how classified standard content can be represented in a machine-executable form without loss of information. The solution consists of an automatic extraction of standard content (here using the example of formulae) and their transfer into a machine-executable knowledge representation form, which can be accessed by different authoring systems. Requirements and design rules can be derived to a higher abstraction level of the “next generation standard” from the knowledge gained during the concrete concept implementation.

**Extraction**

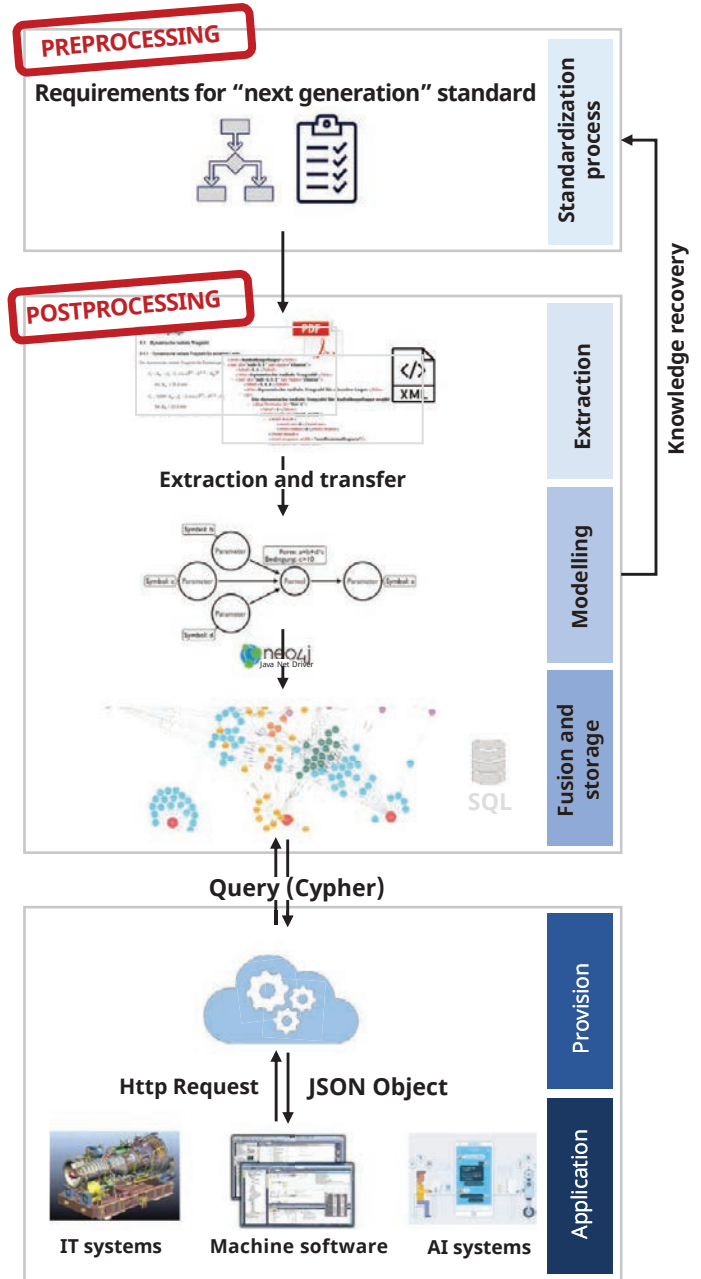
Standards are available to various stakeholders not only as PDF files, but also in XML format, where standard elements such as formulae, tables, diagrams and/or text are tagged in the source code. Extraction describes the actual process of reading out relevant information. For this purpose, an XML parser can be used, which recognizes the marked (formula) elements of an XML-based standard and transforms them into a predefined graph pattern.

**Modelling**

Modelling allows a simple and clear knowledge representation under consideration of the machine executability. The goal is to achieve an automated transfer of extracted information (1:1). Standardized graph patterns are defined according to the kind of the standard element. Using the example of formulae it becomes obvious that parameters as well as operators can be modelled as nodes and their relations as edges.

**Fusion/storage**

The procedural step “Fusion and storage” describes the possibility to aggregate all separately generated graph patterns (for formulae, tables, diagrams, texts) to an extendable knowledge net in a database. This enables on the one hand the elimination of all redundant nodes and, on the other hand, the restoration of relationships between individual standard elements.



**Figure 36:** Bottom-up approach to post-processing of standards

**Provision**

The provision of information in the bottom-up approach serves to decouple the source of knowledge (here: graph database) and its use (here: user program). Via a web service, which can be accessed by different authoring systems, queries in the built up graph database are executed and returned.

**Application**

Application here means the use of digitized standard content. The number of possible applications, such as in the area of IT systems (e.g. CAD), machine software or AI-based application systems is extensive. Based on the request of an authoring system, the relevant information is identified in the database and transferred to the authoring system.

Ultimately, the bottom-up approach allows conclusions to be drawn for the standardization process. This in turn enables the elimination of manual, error-prone process steps, significant time savings in the transfer of standard content into company processes, an increase in quality by ensuring the continuous traceability of standard content, as well as a lower adjustment effort for updates of standards

The post-processing of the existing, very large pool of standards is reaching its capacity limits and would only be economically viable for defined subject areas. The use of artificial intelligence in the extraction phase of the bottom-up approach is to be investigated in order to support this step using machines.

**11.4.3 Top-down method – Development of SMART standards**

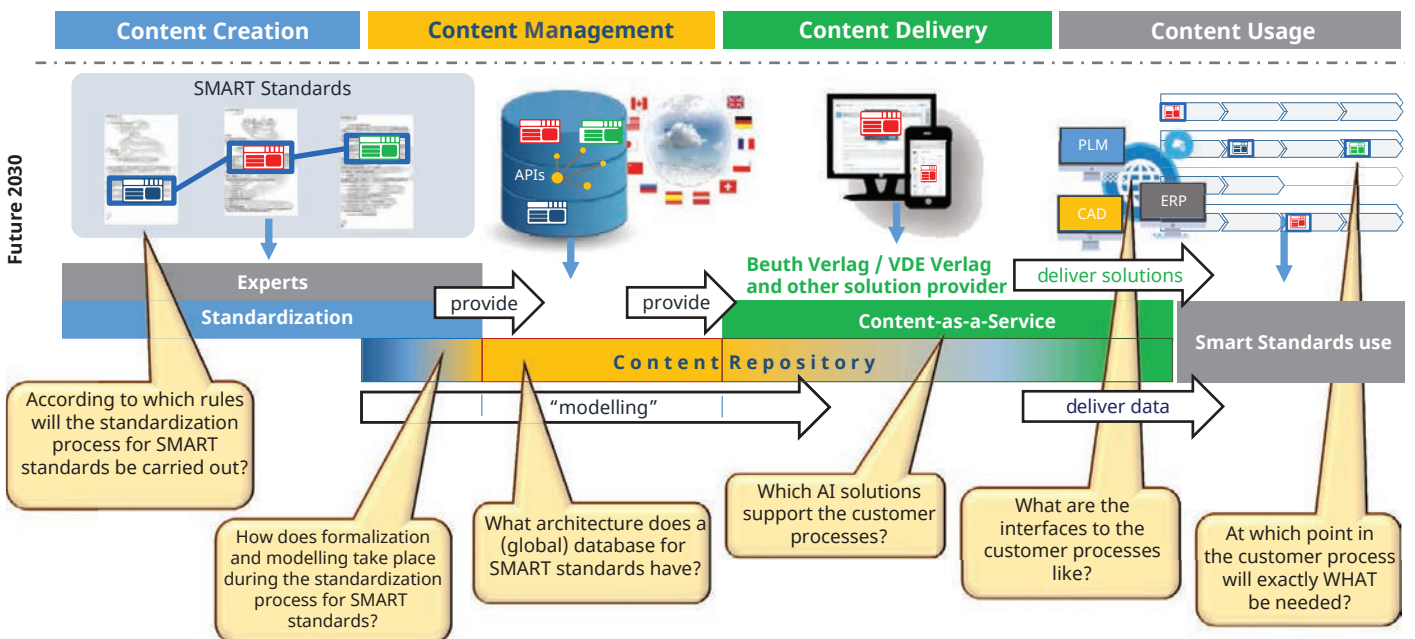
Promising concepts are currently being developed by DIN/DKE, CEN/CENELEC and ISO/IEC to develop appropriate forms of presentation already during the standardization process (preprocessing, “top-down approach”), which allow the conversion into machine-interpretable formats [311], [258].

The overall process for Level 4 partly requires integrated overarching action by those responsible for the process, so that previous responsibility boundaries (Levels 1 to 3) must be reconsidered and redefined. The content responsibility for “content creation” must definitely be located in the process of developing the standards – the primary content. There is no longer a need for postprocessing in the sense of a subsequent interpretation of content for further processing.

**Level 4**

The main questions to be answered are listed in Figure 37. With the answering of the questions, new ground is partially broken, especially in the interaction with AI-based application processes. A simplified representation visualizes the target image of an overall process as subfunctions SM|ART|KI, see Figure 38:

**Level 4**



**Figure 37:** Level 4 process and relevant questions

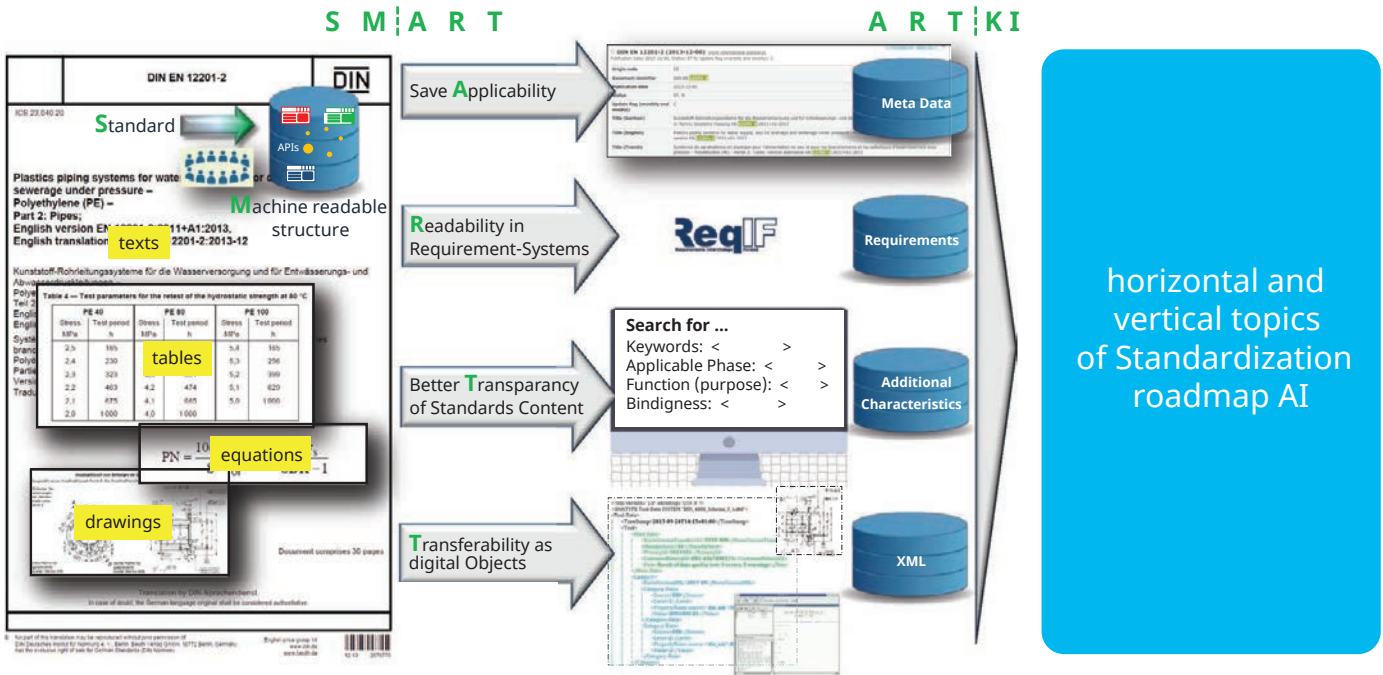


Figure 38: Be SMART – Standards output in a practical application context

SM – Transformation function (Content Creation):  
Standards and specifications are machine-interpretable models

ART – Transport function (Content Management & Delivery):  
Machine-interpretable information as delivery models

AI – Use function (Content Usage):  
Delivery forms as AI-based application processes

As stated in the present AI Roadmap in Chapter 1, “the economic fields of application for AI are extremely diverse”. An addition: They are infinitely diverse. AI is relevant for almost all sectors of the economy, and also for other areas of application outside the economy, and is found both in the form of components in end products and services and in the productive core processes and support processes within companies. Thus, it is clear that an overall process cannot be built from the use cases.

The following is a summary of the main methodological framework in which some of the questions were developed and in order to achieve the goal [309]. They concern the fulfilment of the transport function (see above):

- the development process of SMART standards and
- the content structure (information model).

### Development process of SMART standards

The demand for the granulation of standards down to the smallest useful information elements and their marking is one of the core tasks to be solved. The methodical development by experts of different disciplines, as well as the development of defined and delimitable partial results of standardization in comprehensible steps are an essential success factor. The following describes a procedure that can contribute to the success of the objective.

The current standardization process aims to publish an agreed and tested standard as the final result of a standardization project. In terms of SMART Standards, what will be the work result to be delivered in the future? One will no longer be able to produce only (but necessarily also) the “one final” work result. The various forms of presentation of the subject of the standard must be documented for product liability reasons and to ensure that the partial results are transparent and traceable. The development of SMART standards is de facto a development process, comparable to the systematics of product development processes, for which the various development responsibilities must be presented in a similarly transparent manner.

The methodical procedure according to VDI 2221 [312], [313] provides a basis for describing future standardization



processes (phases, work steps, work results, actors). In the current preparatory work at DIN/DKE and CCMC, the systematic modelling of a standardization object is being tested in pilot projects and consolidated in further projects.

An approach for this is summarized in the following. The **phases**, **work results**, and “actors” are indicated by the chosen formatting.

**Standard proposal phase**

Before the concrete processing of a standardization project, the relevance and financing are checked, the standards committee is assigned and the stakeholders are identified. In the future, the implementation level of the SMART standards solution to be developed will have to be determined by an extended “circle of decision-makers”, in which the “standards user” now also has an important input to contribute. The implementation stage determines the degree of digitization of a standardization project:

a) A standard is created according to the current creation and usage scenario (see Figure 26), with delivery of the entire content in XML, e.g. for further use in editorial systems or other usage environments. Additionally, digital objects defined in

the standard (e.g. tables, formulae, graphics) can be specified for direct use in customer systems.

b) The goal is to develop a SMART standard for a future creation and use scenario (see Figure 37) in the H2H and H2M forms of presentation for the direct AI-based use of granular standard contents in customer processes (M2M), see Figure 39.

**I. Requirements definition phase**

The requirements for the subject of standardization must – as before – be formally defined and documented in accordance with DIN 820 [314] and as regards content by the “technical experts in the standards committees” (hereinafter “technical experts”). Furthermore, it is necessary to describe the desired partial results of the subsequent phases, depending on the decided implementation stage (= degree of digitization).

**II. Standardization phase (concept)**

The language (prose) of the “technical experts” is currently and even in the foreseeable future not suitable for directly transforming it into a machine-interpretable form in terms of

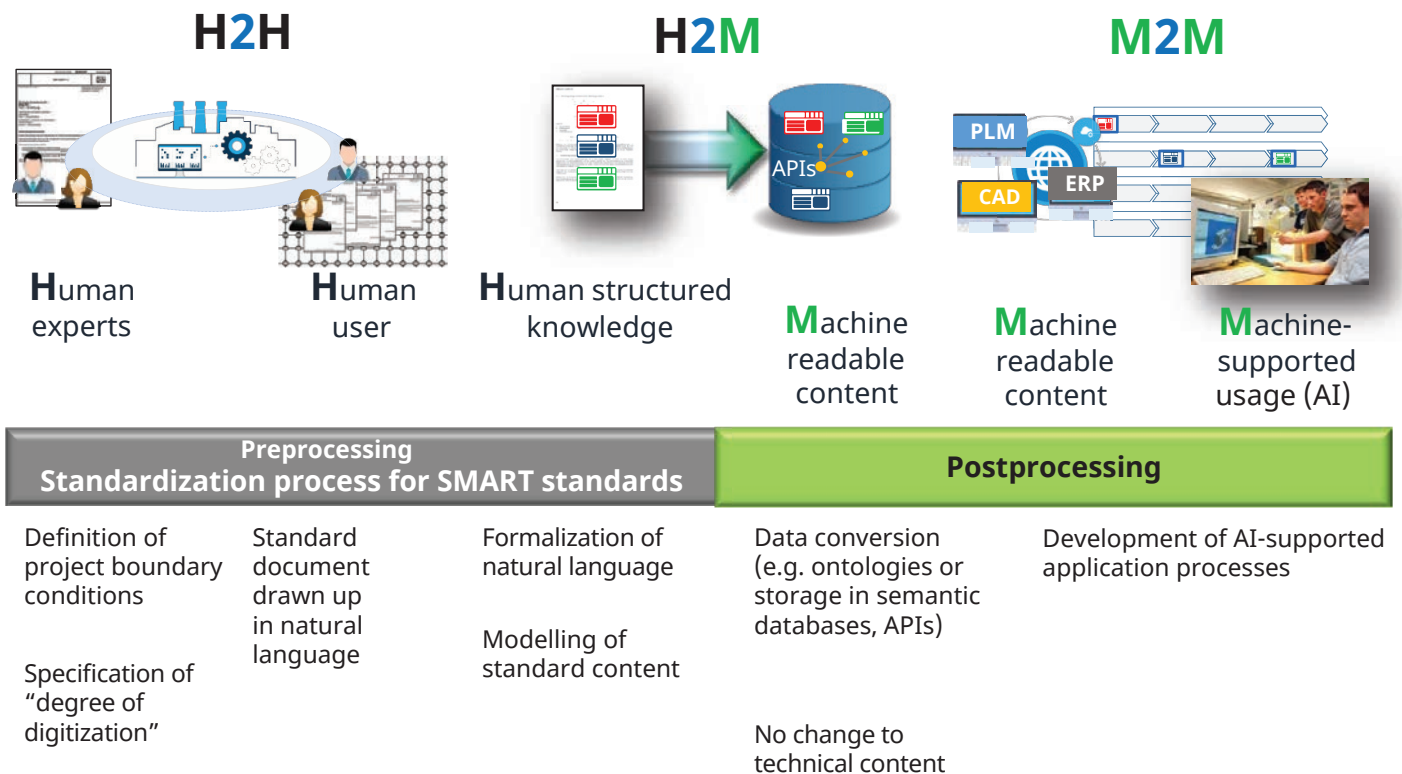


Figure 39: Development of SMART standards for downstream application processes

SMART standards. This natural-language form of presentation is, however, necessary in order to articulate and consolidate expert knowledge at all, and to coordinate it in accordance with DIN 820 [314] and other regulations.

Following a methodical approach, it will therefore be necessary that such formulations are preserved and **documented** as an interim result of phase II. of a standardization project (product liability) and thus represent the traceable input for the subsequent steps.

### III. Formalization phase (Development)

To create a **formal form of presentation** from II. must remain the task of the “technical experts”. These experts must have “extended structuring competencies” in order to carry out this part of the standardization process according to rules that are to be defined further.

Currently, a target-oriented concept is being pursued: The implementation is to be based on the “Semantic Triple”, which has been designed, for example, within the Resource Description Framework (RDF) as an elementary building block for the Semantic Web, a global network of information (see in phase IV.).

The application and thus implementation of the rules can be described by “provisions” and can be ensured or supported by tools to be developed (e.g. XML-based), which achieve or promote formalization.

The tables or other forms of presentation are an interim result of phase III., which must be documented (product liability, traceability). It represents the input for the next step.

### IV. Modelling phase

In the last standardization step of a standardization project, the “technical experts with IT modelling competencies” must become active. Starting from the unambiguous representation form III, these “modelling experts” can build **models, such as triplestores, DB structures, etc.**

The triple represents the link between the natural language and the data structure: A subject-predicate-object relationship is used to define a statement from known elements, e.g. “grass is green”, which itself can be used as a subject or object. Such nesting allows the formulation of complex sentences or specifications in human language, an IT-side control during content entry ensures that the (nested) triple structure is adhered to, and at the same time generates machine-readable content.

Competence in the content of standards for the transformation process is still conducive to identifying inconsistencies in the content of the transfer from phase III. to IV. If an unambiguous transfer to corresponding models is not possible for individual formal representations, this must be reworked in phase III.

In the sense of a consistent methodology, the results of phase IV. must be carefully and comprehensively **documented** (product liability, traceability). They represent the input for further implementations (see e.g. [315]) in the application context.

It should be noted that although the four phases will be worked through one after the other, it will always be an iterative process. It can also be assumed that integrated processing of phases II. and III. or even II. to IV. will be possible for standardization subjects whose forms of presentation are designed from the outset to be “IT implementation-oriented”. In very few sub-areas of standardization, e.g. material properties, STLB construction standardization, this is already possible today.

II. Standardization phase: Currently, the language (prose) of technical experts cannot be directly transformed into 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.

III. Formalization and IV. Modelling: Transformation using “semantic triples” can provide a direct interface to AI processes. Close cooperation is required.

### Required content structures – the information model

Specifications in standards (requirements, recommendations, etc.) usually consist of recurring elements that are linked together according to a certain pattern. For example, these often describe a system or a function in connection with a certain performance or property, which may only be maintained under certain conditions.

An information model suitable for this purpose contains a template for the formulation of specifications that can be used as generally as possible and defines the modelling of the elements contained therein according to the triple concept. Further, the information model contains metadata to

the definitions, by which a subsequent treatment is facilitated (e.g. obligation or function of the definition). Ideally, these metadata can be clearly derived from the standard content or project data and then belong to the primary content. An

excerpt of the information model is shown in Table 15. In the following some substantial characteristics of the information model are described.

**Table 15:** The Information model for SMART standards (excerpt and status as of July 2020, DIN e.V.)

No.	property	value	occurrence	data type	definition
<b>Elements forming a provision or requirement topic</b>					
0	title	text	optional	content	heading or title (subject + action)
1	system   subject	~	required	content	subject; product; system
2	subject-type	~	optional	metadata	connection (link)
2.1		• system	~	value	defined description of the system
2.2		• term	~	value	defined description of the system
3.1	action   modal verb	~	required	content	bindingness word; modal verb; auxiliary verb
3.2	action   main verb	~	required	content	main verb; strong verb; full verb; action
4	actor	~	required	content	effective site; inherited from scope (document) or (sub-)clause
5	performance   object	~	required	content	object; performance
6	object-type	~	optional	metadata	connection (link)
6.1		• term	~	value	defined description of the performance
6.2		• provision	~	value	provision of the defined types
6.3		• numeric value	~	value	numeric value
6.4		• unit	~	value	unit
7	condition	~	optional	content	conditions
8	margin	~	optional	content	deviations; limits; tolerances

No.	property	value	occurrence	data type	definition
<b>Attributes to the provision or requirement topic</b>					
8	relation	relation	required		
8.1		• and	~	value	A1 and A2
8.10		• xor	~	value	either A1 or A2
9	bindingness	~	required	metadata	degree of compulsion
9.1		• capability	~	value	capability
9.5		• requirement	~	value	requirement
10	type		required	metadata	interaction of the provision or requirement
10.1		• activity	~	value	activity (process)
10.6		• verification	~	value	verification (method of proof)
11	function	~	required	metadata	requirement or standard function
11.1		• availability	≈	value	
11.11		• sustainability	≈	value	
12	smart-tag		optional	metadata	allow marking which SMART property is equivalent for this triple
12.1		• actor	≈	value	effective site; inherited from scope (document) or (sub-)clause
12.5		• system	≈	value	subject; product; system

No.	property	value	occurrence	data type	definition
13	classification	~	~	metadata	classification [inherited from document or topic]
14	date of activation	YYYY((-MM)?-DD)?	required	metadata	date of publication (dop) [inherited from document or topic]
15	date of creation	YYYY((-MM)?-DD)?	required	metadata	date of availability (doa) [inherited from document or topic]
16	date of deactivation	YYYY((-MM)?-DD)?	required	metadata	date of withdrawal (dow) [inherited from document or topic]
17	date of last change	YYYY((-MM)?-DD)?	required	metadata	date of availability (doa) [inherited from document or topic]
18	date of revision	YYYY((-MM)?-DD)?	required	metadata	date of publication (dop) [inherited from document or topic]
19	date of version	YYYY((-MM)?-DD)?	required	metadata	date of publication (dop) [inherited from document or topic]
20	guid	~	required	metadata	global unique identifier
21	informative	text	optional	metadata	system of interest, rationale, explanation, examples
22	keywords	~	optional	metadata	discriptors [inherited from document or topic]
23	language	ISO 639-1	required	metadata	2-letter language code in lower case
24	source-of	~	required	metadata	source; reference to standard or law text
25	status	~	required	metadata	stage-code [inherited from document or topic]
26	version number	YYYY-MM-DD hh:mm:ss	required	metadata	time stamp of last change of requirement or date of publication

**The standardization function as the main structuring feature:** The basic idea of earlier considerations was to combine contents with the same function in order to successively build a system of networked modules. For the current task – the development of standards – it is analogous to structure the contents to be developed according to characteristics to be defined in such a way that integration into an increasingly growing SMART standards environment is possible. The standardization function therefore has an important meaning for the current task.

**Definition “standardization function“:** “A standardized element (smallest meaningful ‘standard granulate’, e.g. sentence, clause, formula, data, figure etc.) is only formulated with a defined bindingness in a standard if a purpose is to be fulfilled thereby”.

This purpose can fulfil various intended subfunctions:

- Communication: Create or promote understanding, enable communication through uniform terminology.
- Quality: Specify requirements, safety measures and sustainability processes.
- Testing: Specify conditions, procedures and evaluations.
- Safety: Specify requirements by means of characteristics for tangible (e.g. consumer goods) and intangible (e.g. services) objects.
- Fundamentals: Define uniform standards of action.
- Accumulation: Reduce the variety of tangible and intangible objects; simplify procedures (processes) and reduce expenses (time, costs, material).
- Recycling: Regulate the recycling/reuse and reuse/utilization of resources.
- Relationships: Identify and align normative relationships.
- Interoperability: Enable exchange of tangible and intangible assets; promote technology, movement of goods, enable applications.

The development and specification of further standardization functions is not yet complete.

**Further standardization-related features for structuring:** The formal presentation of the different subjects of standardization will not be a hurdle in practice as long as the structuring features are understandable and traceable.

In addition to the main feature “standardization function” described above, further structuring features and their characteristics should clearly represent the individual subject of standardization. These are (at present):

- Technical characteristics of a standardization subject:
  - The presentation can be made in a table. Currently, the CCMC projects follow a linguistic approach (Subject, Action, Object). An evaluation of the user acceptance of a suitable form of presentation is not yet available.
  - Binding characteristics according to DIN 820 [314]:
    - obligation (“shall”)
    - recommendation (“should”)
    - permission (“may”)
    - possibility (“can”)
  - Interaction of the subject of standardization, according to the definition “Function of the Objective” from the preliminary considerations in the CCMC pilot projects [316]:
    - Activity
    - Constraint
    - Integral aspect
    - Interface
    - Verification
  - Metadata of the entire document, according to the state of the indexing methodology and the requirements of a standards management [317]. The extent to which the findings from the use of semantic methods can be integrated should be further examined.

**Application-related characteristics:** Process application aspects (“Who uses which standardization content?”) are generally not defined in the standardization process. A meaningful exception should be the characteristic “place of action” – an additional piece of information which has not been consistently found in standards to date.

**Definition “Place of action“:** “The place where the standardized event takes effect marks the place of action.”

For example:

- Action in the process: e.g. construction (for example in the concept or elaboration), after sales, reproduction etc.
- Action in the case of functions to be fulfilled: e.g. connecting

Example for future development processes using the information model (see Figure 40)

**I. Requirements definition phase**

The degree of digitization is determined in the standards proposal phase: A standardization project corresponding to phases I. to IV. is to be developed as a SMART standard. The requirements for this are documented.

**II. Standardization phase**

The requirements are formulated by the experts in textual form, for example:

“If the temperature is above 50°C or the pressure exceeds 50 MPa, the tube and fittings shall either be made of material conforming to EN 1234 or have a modulus of elasticity between 15,000 N/mm<sup>2</sup> and 18,000 N/mm<sup>2</sup>.”

→ Suitable for existing (“word-based”) standardization processes (e.g. basis for draft voting)

**III. Formalization phase**

The semantic triple is the basis for formalization:

- It ensures that specifications are only made for defined elements
- It enables the structuring of standard contents on the basis of unique information elements
- Nesting of triplets is possible.
- **Triple structuring increases the quality of documents in Level 3 considerably and sets the foundation for Level 4**

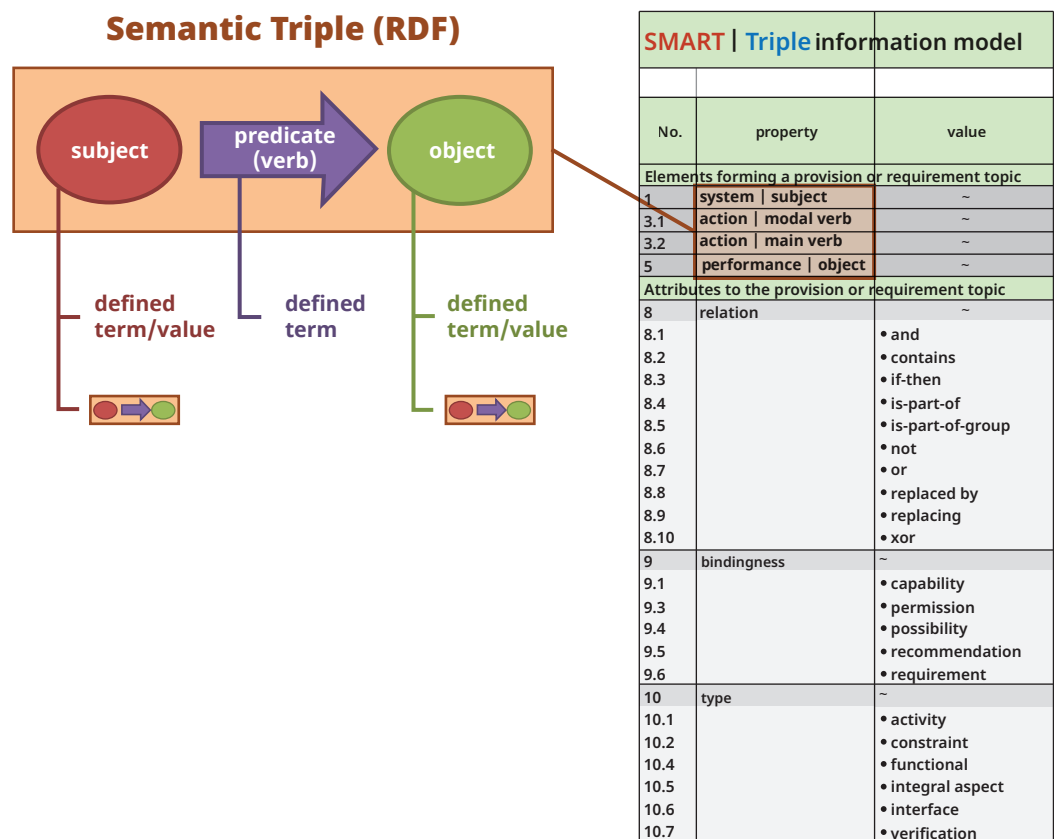
**IV. Modelling phase**

As a tool for implementing the concept, DIN is currently testing specially developed graphical user interfaces that allow the standards committee to concentrate on the plain text of the standard and generate the described data structures in the background (see Figure 41).

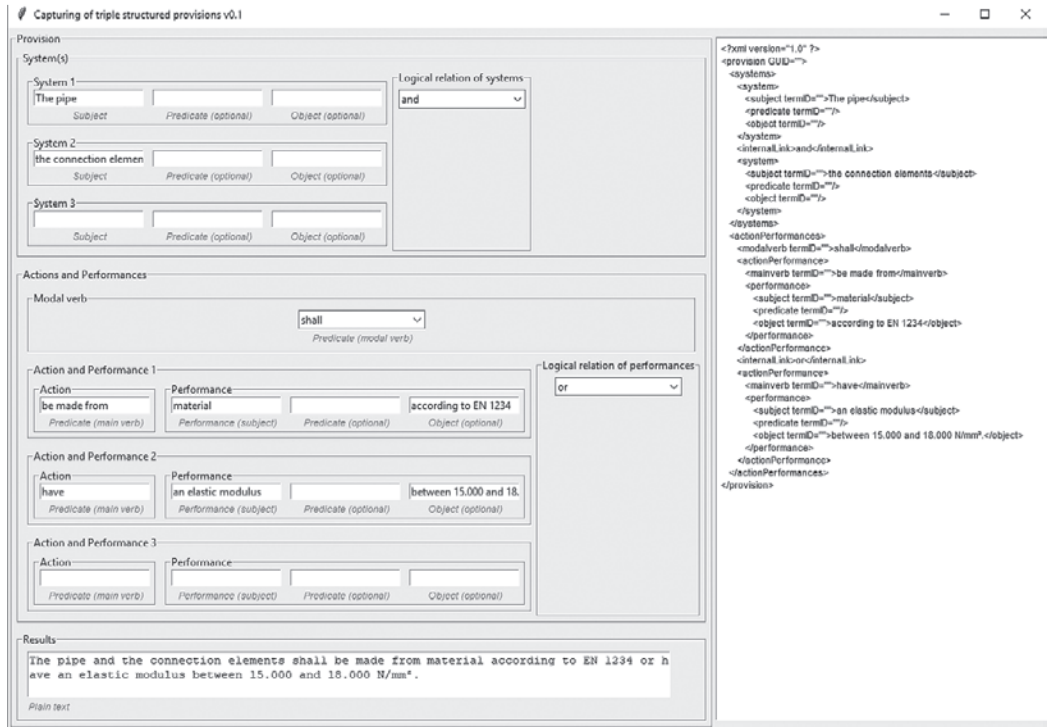
**Experimental GUI for drawing up requirements**

The metadata of interest to a requirements management system, for example, can then be generated from these data structures in a clear and automated manner. Thus, for ex-

Figure 40: Triple structuring for future development processes using the information model



**Figure 41:** Graphical user interface as a tool for implementing the concept



ample, the binding nature of a specification results from the modal verb, which can be uniquely identified (through triple structuring), or the respective object of standardization can be achieved by evaluating the linked subject elements.

**EXAMPLES FOR POSTPROCESSING**

XML result from GUI (see Figure 42)

→ **The natural language remains. All semantic information is stored in XML and can be used for Level 4.**

OWL/RDF translation (see Figure 43)

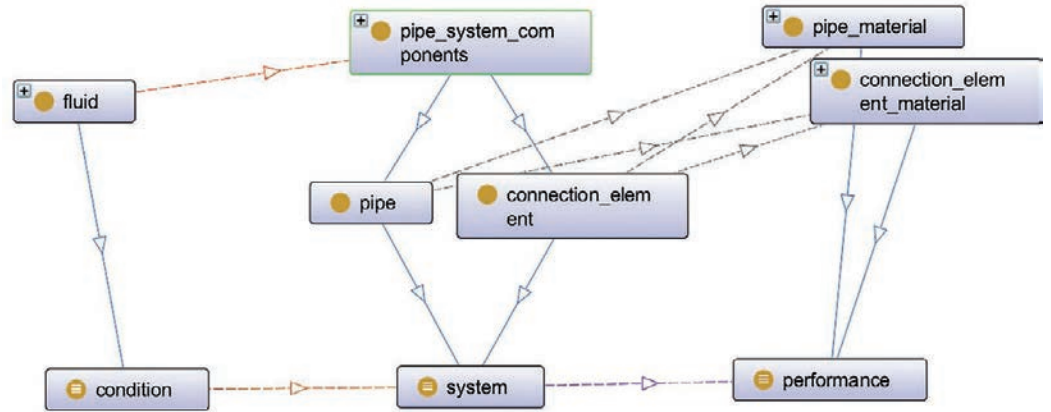
But also the automated generation of ontologies (e.g. OWL according to W3C) for the secure application of AI systems is possible. This covers a large number of conceivable applications for SMART standards.

→ **Translation into OWL/RDF can be done in a downstream automated process.**

**Figure 42:** Example data structure of the metadata of interest of the request system

```
<?xml version="1.0" encoding="UTF-8"?>
<req bindingness="requirement" reqtype="integral aspect" guid="8d53fela-0ca6-4e29-8327-fdb684cdb82c">
  <relations></relations>
  <functions></functions>
  <req-condition elements="2" internal-link="or" external-link="if">
    <external-link term-id="882c695b-eb80-46ae-9197-0df5080721d6">
      if
    </external-link>
    <element type="triple" number="1">
      <triple>
        <subject type="term" term-id="9d90ed81-9530-4a4c-b7c6-2d3988dc8c0f">
          temperature
        </subject>
        <predicate type="term" term-id="9eefe750-eald-4e16-a5da-d42fd417c527">
          is
        </predicate>
        <object type="value" term-id="">
          above 50°C
        </object>
      </triple>
    </element>
    <internal-link term-id="38c68a7d-0964-43cc-89e3-efb1385896ae">
      or
    </internal-link>
  </req-condition>
</req>
```

**Figure 43:** Automatic translation in OWL/RDF











**DIN e.V.**

Burggrafenstr. 6

10787 Berlin

Tel: +49 30 2601-0

Email: [presse@din.de](mailto:presse@din.de)

Website: [www.din.de](http://www.din.de)



**DKE German Commission for Electrical,  
Electronic & Information Technologies of DIN and VDE**

Stresemannallee 15

60596 Frankfurt am Main

Tel: +49 69 6308-0

Fax: +49 69 08-9863

Email: [standardisierung@vde.com](mailto:standardisierung@vde.com)

Website: [www.dke.de](http://www.dke.de)

November 2020