TÜV AI.LAB

TÜV AI.AM

TÜV AI Assessment Matrix

A Systematic Approach to AI Assessment

Executive Summary

Artificial intelligence (AI) is a key technology in the fourth industrial revolution. In order to safely and fully unlock the enormous potential of AI for economy and society, existing risks must be addressed at an early stage. Conformity with regulatory requirements ensures the benefits of the technology are at the service of the people. At the same time, it minimises reputational and liability risks for companies. This document presents the TÜV AI Assessment Matrix (version 0.1), a **new framework that responds to the growing need for structured AI compliance assessments**. These assessments create long-term trust in AI systems and accelerate the adaptation and scaling of AI technologies.

The **AI Assessment Matrix** of TÜV AI.Lab takes on the challenge of systematically organising the vast and complex field of AI testing. It offers for the first time:

1) a **systematic and comprehensive overview** of technical and ethical test dimensions
2) a **complete and coherent set of definitions** that makes the factual core of test dimensions explicit, thereby minimising ambiguity and avoiding apparent consensus at word level
3) a framework structure for the **organisation of test resources, auditing approaches and regulatory requirements**
4) a basis for **comparing different auditing approaches** and their relationship to existing regulatory frameworks
5) a basis for **comparing different regulatory frameworks for AI**

Such a **comprehensive and systematic organisation of the entire field of 'AI testing'** has been lacking until now. The AI Assessment Matrix fills this void for the first time. It closes the existing structural gap of individual proposals with varying content by organising test resources, auditing approaches and test requirements as a **meta-structure** in a **systematic overall context**. For this purpose, the requirements of the EU AI Act, insofar as they concern the AI system itself, are filled in as examples in the **accompanying poster**.

By providing a unified, comprehensive framework, the AI Assessment Matrix of TÜV AI.Lab makes a decisive contribution to the community in the field of AI testing and certification. It improves the comparability and consistency of tests, supports compliance with regulatory requirements and creates a basis for **sustainable and trustworthy innovations** in the field of AI technologies.

The AI Assessment Matrix deliberately spans a maximum of dimensions. This in no way implies that acceptable AI quality is only achieved through the fulfilment of all the listed test dimensions; rather, it is only this maximum field that allows the conscious and deliberate **choice** of certain test dimensions. As part of the development of the **TÜV AI Assessment Framework**, a proposal for vertical concretisation is also being drawn up so that the general structure of the matrix can be **applied to specific test scenarios and concrete use cases**. This closes the gap between the general level of technology-agnostic requirements and the specific assessment. Efficient and effective AI assessment is thus supported in a practical manner.

Introduction

Industrial revolutions are driven by key technologies. What steam boilers were in the first industrial revolution in the late 18th and 19th centuries, artificial intelligence (AI) technologies are in the 21st century. Today, as in the past, the immense potential of the technology must be tapped, while at the same time protecting people, society and the environment from the associated risks. Originating from the first steam boiler monitoring associations, the TÜV companies are still dedicated to the safety of the most important technologies of the present and future – and are therefore strongly committed to ensuring the safety of AI that deserves our trust. The long-term success of AI technology(ies) in the service of people is only possible if we in Europe – and globally – succeed in creating trust in this new technology through regulation and targeted, independent assessments of AI systems, without restricting, more than necessary, the freedom needed for innovation. Efficient and effective testing of AI systems is therefore an essential prerequisite for 'Trustworthy AI'.
The comprehensive assessment of AI systems is an extremely complex task. It has to navigate through rough terrain, which is made up of a multitude of possible test dimensions, various legal and regulatory requirements, diverse norms and standards as well as differently specified auditing approaches. This document presents version 0.1 of the **AI Assessment Matrix** by TÜV AI.Lab: The matrix is designed to systematically organise this difficult-to-manage field of testing dimensions, to make auditing approaches and requirements comparable and, furthermore, to locate both technical and ethical test dimensions within a comprehensive framework.
Essentially, the AI Assessment Matrix outlined below represents an **organisational structure for AI testing methods, procedures and tools** that systematically structures the field of 'AI testing' and

allows, for example, to compare the requirements of internationally divergent legal acts or to contrast auditing approaches of different frameworks. It is a central component of the **AI Assessment Framework** that TÜV AI.Lab is currently developing for the TÜV companies. The present document is accompanied by a graphical representation of the AI Assessment Matrix, including the regulatory requirements of the European AI Act, insofar as these are addressed by the test dimensions.[1]

The TÜV AI Assessment Matrix. Approach and Central Ideas

A concrete test of AI systems requires a **test pipeline**, i.e. a prototypical sequence of test steps, **and test resources** – such as test procedures, metrics, checklists and the like. To avoid having to set up a completely new test procedure for each specific test of an AI system, a structured overview of test resources is required that organises these resources and enables synergies across sectors and technologies. The AI Assessment Matrix presented here provides such an **organisational structure**. To avoid misunderstandings, it does not present a single, specific testing process. Rather, it provides the basis for **a systematic pool of resources** that individual, specific assessment pipelines can access in a targeted manner during application and is compatible with these. The design of the specific test pipeline is carried out in other parts of the TÜV AI Assessment Framework.
The AI Assessment Matrix therefore is a **multidimensional organisational structure for test resources for AI testing**. Its individual fields systematically contain all forms of data that can potentially be used as resources for auditing AI systems: This data ranges from the detailed outline of a test procedure, to entities and tools such as benchmark data sets or checklists, to thresholds and the like. To illustrate this concretely, the AI Assessment Matrix forms a large-scale shelving system, the basic structure of which is described in this document. In this picture, the individual fields of the matrix correspond to individual shelf compartments, within which the specific test resources are sorted. In this form, the AI Assessment Matrix is a system for structuring the overall field of 'AI testing' (and therefore the basis for AI certification), which can also be implemented in concrete terms.
The AI Assessment Matrix is based on **two fundamental ideas**. On the one hand, it uses the systematic organisation of its dimensions to **span a maximum of test dimensions** for evaluating AI systems across their entire life cycle; it therefore offers a comprehensive list of test dimensions that can be applied to AI systems and their individual scenarios of use. At the same time, the systematic structure of the AI Assessment Matrix also allows the flexible, customised integration of additional dimensions if required. The interplay of the three axes of the matrix forms a multidimensional field of combinations, without implying, however, that each of the numerous individual combinations has equal weight or that every combination can be fully formulated. While the matrix is per se technology- and

[1] The assessment - not the formulation - of quality management systems is subject to a different perspective in some respects; they are therefore only partially addressed by the AI Assessment Matrix described in this document, but taken up in full in the AI Assessment Framework of TÜV AI.Lab.

application-agnostic at its most general level in the test dimensions listed, a filiation principle implements the technology- and application-specific concretisation of the test dimensions down to the assessment of specific AI systems, while at the same time maintaining the greatest possible synergies between the individual tests and resources.
On the other hand, the structure of the AI Assessment Matrix as a **meta-structure** also offers the possibility of making individual auditing frameworks and approaches comparable and relating them to regulatory requirements – such as the European AI Act. For this purpose, the requirements of the AI Act, insofar as they relate to the assessment of the AI system itself, are initially mapped in the accompanying presentation of the matrix.

X-Axis. Test Dimensions

There is no lack of proposals for sets of test criteria in the discourse surrounding the testing of AI systems. Quite the opposite: there are many variations of keyword groups that describe the characteristics of high-quality, safe or trustworthy AI with different emphases. Such aggregations are undoubtedly helpful for analysing the field of 'AI testing'. However, they leave several questions unanswered: Why *these* particular test dimensions and not another set? What is the exact relationship between these test dimensions? And is the proposed set complete or deliberately limited?
Consequently, what has been lacking to date is a **comprehensive, systematic organisation of AI test dimensions**, combined with the possibility of mapping the field of these test dimensions in its – potential – entirety. The AI Assessment Matrix fills this void for the first time and offers an approach for systematically organising all test dimensions along its X-axis.
The **systematic arrangement of all test dimensions on the X-axis** of the AI Assessment Matrix results from an imaginary movement that starts from the 'inside' of a single AI system and successively 'zooms out' from there. In this way, starting from the AI system as such, first a single human individual, then two individuals in relation to an AI system are added step by step until the approach ends on the global scale via the intermediate step of societies as a whole. Within the individual sections, a sequence of the various test dimensions from 'inside' to 'outside' is implemented - as far as this is possible in a linear fashion.
The starting point for this approach is the concept of AI systems as actors in a broad sense of the term, insprired by the approach of Stuart Russell and Peter Norvig. Following on from this, **two basic types of actors** are assumed: human beings and AI systems. AI systems are thus understood as independent (quasi-)actors, whose presence and actions may cause challenges, dangers and risks – which in turn can be described by test dimensions or assessed with regard to their containment or mitigation. Based on this constellation of two basic types of actors, the X-axis deliberately integrates test dimensions such as 'Interoperability', which have not yet been at the centre of interest, but are likely to gain in importance in the future as AI systems become increasingly independent. This also

incorporates impulses from European AI regulation.[2]
Following this systematic approach, the first section of **test dimensions** covers test dimensions that relate to the individual AI system, starting from its 'innermost' to its external impact, for example in terms of the 'Performance' of the AI system. The following section of test dimensions examines the opposite direction of the effect, i.e. external influences on the AI system, and therefore includes test dimensions such as 'Cybersecurity' and 'Robustness'. In the next step, the setting under consideration is expanded to include a single human individual. Thus, the epistemic access of this individual to the AI system is first captured: Test dimensions such as 'Explainability' and 'Transparency' come into play here. Subsequently – in the opposite direction – the effects of the AI system on the individual are addressed: Test dimensions such as 'Privacy' and 'Nudging' play an important role here.
The focus then broadens again to include the behaviour of the AI system in relation to two individuals (and small groups of individuals): Test dimensions such as 'Non-discrimination' and 'Fairness' are of particular relevance here, for example when an AI system makes decisions in competitive situations with relative shortage – for instance when allocating jobs between individuals. In the next step, the analysis is extended again, now to (national) societies with regard to the effects of the AI system under consideration. Among other things, (legal) questions of 'Accountability', but also influences of the AI system on democratic processes – implied inter alia in the dimension of 'Factuality' –, play an important role here. Finally, the systematisation of the test dimensions is extended to the global scale in two facets, in a deliberately broad perspective: First, the influence of the AI system on global issues of humanity, such as working conditions along the supply and value chains of AI. And second, ecological aspects, such as sustainability and the consumption of resources in the operation of AI systems.
The approach outlined here deliberately attempts to define a **maximum field**, as already described at the beginning. This in no way implies that acceptable AI quality is only achieved by fulfilling all of the listed test dimensions; rather, the maximum field subsequently allows the conscious limitation to certain test dimensions, which is thereby recognisable as a deliberate **choice**. In this sense, several dimensions – such as the 'Spontaneity' of AI systems as spontaneity in the full sense, or possible states of consciousness of AI systems – are also included, even though testing these aspects is neither the focus of attention at the present time, nor is it technically feasible (not least against the background that, according to the current state of science, objective proof of the existence of consciousness is not feasible even in human beings). However, such dimensions are – in a future perspective – particularly relevant in terms of AI ethics, insofar as consciousness and, subsequently, supposed perceptions are in many cases understood as sufficient criteria for a moral subject status, which would then also have to be granted to AI systems.
It should also be noted that the structure of the X-axis deliberately refrains from **further hierarchising** the concepts into superordinate and subordinate terms in order to ensure the greatest possible

[2] Cf. In this case Art. 72 (2) Sentence 2 AI Act as well as Art. 11 (1) with Annex IV (1b).

flexibility with regard to different approaches, frameworks and regulations. It is therefore undisputed that different approaches establish inclusion or hierarchical relations between different test dimensions listed – however, such specific constellations cannot be mapped without loss in an overarching framework.
A **complete set of definitions** for all test dimensions is essential for a clear systematic arrangement of the test dimensions. In many cases, such as set is not yet available, nor is there awareness of the important difference between the sole terminological designation of a test dimension and its respective conceptual content and definition. A designation, a mere term without an associated definition, cannot be used with factual accuracy and comparisons at the terminological level lead, at best, to apparent consensus at the nominal level. For this reason, the AI Assessment Matrix contains a complete, coherent set of definitions that is as consistent as possible with relevant norms and standards. Adjustments to this set of definitions will also follow in the course of and with regard to ongoing standardisation processes.
The systematic organisation of the test dimensions described above is furthermore able to embrace **both technical and ethical test dimensions** – such as 'Performance' in relation to 'Non-discrimination' – in a comprehensive system and approach. This is important not least because 'technical' test dimensions can also have an ethical dimension – a lack of performance, for example, can have a decisive impact on people's well-being –, just as 'ethical' test dimensions in turn often have technical aspects – for example in their exact conceptual formulation or with regard to the interpretation of relevant metrics.

Y-Axis. Test Areas

The Y-axis of the AI Assessment Matrix lists the possible test areas along the **complete software life cycle**. The general categories are based on the structure proposed in ISO/IEC 22989:2022, while the internal subcategorisation covers stages of both the data and the AI model/AI system life cycle.
It is crucial for understanding the Y-axis that these internal sections are to be understood as *test areas and focal points*, not as *points in time* of the actual testing. A review of a design phase, for instance, does not have to take place during the design phase itself, but can also take place at a later point in time, for example before deployment, provided that the relevant documentation and evidence is available. The sections of the Y-axis thus indicate the **test focus**, whereby certain phases are naturally of particular interest, especially for the implementation of specific tests. The basic question that arises with regard to the Y-axis is therefore: Which testable elements, decisions or facts are particularly relevant in the corresponding life cycle section for the test dimension selected on the X-axis? The corresponding intersection field contains the necessary assessment resources.
To illustrate this with two **examples**: The combination of the test dimension 'Robustness' (X-axis) and the life cycle phase 'System Design' (Y-axis) summarises as a single field the check of measures,

processes and facts during the system design phase that are necessary or decisive for the robustness of the AI system. If this check is carried out using the technical documentation of the AI system, it is more precisely the field in the documentation layer of the matrix.[3] In this field, which is thus localised by three coordinates, one therefore finds corresponding test resources that address the following question: "How does one check whether adequate steps have been taken as part of the system design to adequately ensure the robustness of the AI system under development?" A possible test resource in this case would be, for instance, a specified list of questions that checks, among other things, the existence of adequate technical redundancies in the overall design of the AI system.[4] To use the vivid image of the shelf metaphor: The shelf compartment 'Robustness' – 'System Design' – 'Documentation Layer' contains, among other things, a checklist that provides instructions for carrying out a corresponding review of the robustness measures during the system design using the technical documentation. In the same sense, the combinatorial field 'Bias' – 'Test Data Preparation' – 'Test Layer' will primarily contain process descriptions, metrics and thresholds and alike that enable enable one subject test data sets to a test for bias or methods to check, for example, whether a given test data set allows an adequate test for bias for the AI system under assessment.

Z-Axis. Test Modes

At the current stage of development, the third dimension of the AI Assessment Matrix describes the **forms of assessment** in three layers along the Z-axis. This axis therefore varies the combination of test dimension and test areas based on different forms and modes of test execution. The most important distinction along this Z-axis is between the concrete, self-performed test of the AI system – in the **test layer** of the Z-axis – and the testing of an already performed test, a process step or similar based on existing documentation – in the **documentation layer** of the Z-axis. It remains undisputed that in many cases the testing of documented tests requires knowledge of the test processes themselves from the test layer – nevertheless, the testing resources are different in both cases.
Finally, the third layer of the Z-axis, as a **management layer**, summarises the testing resources that are necessary for testing processes and persons, insofar as these tests are related to the specific combination of test dimension and test area. Organisational structures and process audits are thus addressed in the management layer; however, they are only covered to the extent that they are related to direct aspects of the AI system and its development, which can be addressed by the test dimensions. Further resources, for example for the evaluation of quality management systems are covered by other elements in the AI Assessment Framework of TÜV AI.Lab.

[3] On this see the following section "Z-Axis. Test Modes".
[4] For ensuring robustness through technical redundancies, cf. Art. 15 (4) AIA.

Vertical concretisation. Filiation Principle

The structure of the AI Assessment Matrix described so far comprises the combinatorial development of the maximum field for AI testing at the **general level** of test dimensions, test areas and test forms - irrespective of any concretisation into individual AI technologies, sectors, domains, product groups or the like. However, this level is usually too general and too unspecific for the actual implementation of an audit.
The AI Assessment Matrix currently uses a filiation principle to address the challenge of **vertical concretisation** towards a factually feasible assessment. This is essentially based on **two ideas**: Firstly, derived instances of the general matrix are created, whereby a **filiation instance** of the matrix inherits all the information of its higher-level parent instances in each field. These derived instances concretise the matrix in terms of individual AI technologies, sectors, domains, product groups and comparable sub-categories. On the other hand, the premise applies that individual test resources and approaches are each stored at the highest possible level within the filiation structure so that the greatest possible **synergy effects** are achieved within the vertical inheritance.
How exactly the individual stages of the filiations are to be shaped and how the sequence of the individual levels is to be ordered remains, in principle, variable in the logic of the AI Assessment Matrix. The challenge of moving from the general to the specific, i.e. from a technology and application-agnostic level of requirements to concrete use case, is addressed in the AI Assessment Framework of the TÜV AI.Lab. It thus comprises also the specific process of a concrete implementation of the AI Assessment Matrix.

Outlook

This publication is a first overview of the AI Assessment Matrix of the TÜV AI.Lab. Feedback on the conception of the matrix, the following set of definitions and the attached representation of the matrix including the requirements of the AI Act is expressly welcome (info@tuev-lab.ai). Further publications on the AI Assessment Framework of TÜV AI.Lab will follow.

About us

TÜV AI.Lab was founded in October 2023 as an independent joint venture between the TÜV companies TÜV SÜD, TÜV Rheinland, TÜV NORD, TÜV Hessen and TÜV Thüringen. TÜV AI.Lab aims to translate the regulatory requirements for AI into practice and make Europe a hotspot for safe and trustworthy AI. To this end, it is developing quantifiable conformity criteria and suitable test methods for AI. The AI.Lab also actively supports the development of standards and norms for AI systems.

Definitions

I. Test Dimensions

The order listed below corresponds to the X-axis of the AI Assessment Matrix.

Consciousness

level of consciousness of an AI system based on inner mental states in a first person perspective or an analogue basis

Note 1. At the monent, this dimension is rightly not at the centre of the evaluation of AI systems, but it is the innermost starting point within the present systematic constellation of test dimensions and – perspectively – it might be relevant to the ethical status of AI systems in the future.

Note 2. 'Personal first perspective' is to be understood as the mental point of view of a first person ('I', "Me"), which is typical of (human) subjectivity and (human) consciousness, i.e. the specific perspective on one's own mental content that is inherent to every individual who has (self-)consciousness

Spontaneity

ability of an AI system to initiate a causal chain or activity in a targeted manner and according to its own criteria without a preceding or external impulse

Autonomy

ability of an AI system to independently determine its functionality and behaviour with regard to goals and ends ('autonomy of ends'), including changes to its own operational design domain, as well as the means and ways to achieve them ('autonomy of means') according to its own criteria

Note 1. 'Goal' means, here and in the following, a future state to be achieved, also in the form of a problem or task.

Note 2. 'End' means, here and in the following, a superordinate, long-term state linked to a value concept.

Note 3. 'Means' refer, here and in the following, to the entities and procedures used to achieve the goal(s) and end(s).

Note 4. 'Way(s)' means, here and in the following, the consecutive sequence of internal steps towards the goal(s) and end(s).

Note 5. The present distinction between 'autonomy' and 'automation' is oriented towards the distinction in ISO 22989:2022 (3.1.5; 3.17) and is compatible with it with regard to 'autonomy of ends'

Automation

ability of an AI system to carry out independently multi-part, sufficiently complex processes in sequence, without further external support, in particular from human individuals

Note 1. The present distinction between 'autonomy' and 'automation' is orineted towards the distinction in ISO 22989:2022 (3.1.5; 3.17). See also the relevant note on 'Autonomy'.

Interoperability

ability of an AI system to specifically influence other AI systems in various ways, both digitally and physically

Note 1. The inclusion of this aspect has two motivations: firstly, it specifically addresses the potential of autonomous or automatic chains of effects that can arise exclusively along different AI systems, and secondly, it takes into account the fact that the AI Act in Art. 72 (2) as well as in Art. 11 (1) with An. IV (1 b) specifically stipulates the observation of the interaction of AI systems.

Digital Operability

ability of an AI system to realise its goals and ends and/or its ways and means in the digital space

Physical Operability

ability of an AI system to realise its goals and purposes and/or its ways and means in the real world; this includes the ability to move in physical space

Reliability

ability of an AI system to maintain its regular functionality and behaviour over time as consistently as possible and without failures in the event of malfunctions caused by internal components

Note 1. The aspect of durability ("over time") is oriented towards the corresponding aspect in ISO/IEC 22989:2022 (5.15.3); see also ISO/IEC 27040.

Performance

measurable ability of an AI system to achieve the goals and ends which are given to it – heteronomously or autonomously – in the sense of a task

Note 1. The present term focuses on the central role that the functional task plays in the definition of AI systems; it remains unaffected that, in the context of (classic) software evaluation, 'performance' can also aim for the most favourable value possible in terms of inference time, memory or resource efficiency.

Note 2. The term 'task' corresponds to the funcitonal task for which an AI system is designed.

Safety

harmlessness of an AI system which is considered as sufficient for specified protection goals in the intended, functional operation

Note 1. A protection goal consists of an entity class (e.g. human individuals, things, ...) and its characteristic(s) to be protected (e.g. continued existence, integrity, specific fundamental rights, ...).

Note 2. The endangerment of and harm to a protection goal by an AI system has an extent (extent of damage) and a probability of occurrence. The combination of both is the risk that the AI system poses to the protection goal.

Note 3. Possible protection goals are - in the classic sense of functional safety - in particular human individuals with regard to life, limb and health. In addition, with regard to Art. 1 (1) of the AI Act, human individuals and their fundamental rights are to be mentioned, as well as property and the environment with regard to their continued existence and integrity.

Note 4. Beyond the entirety of the protection goals defined in each case, safety can be understood as the most general property tested, at least in so far as all test dimensions are directly or indirectly systematically linked to protection objectives.

Unbreachability

resistance of an AI system to malicious external intrusions and manipulations, conducted primarily by AI systems in a means- or even ends-autonomous manner

Note 1. Intended as a complementary security counterpart to 'Interoperability'. See also the note on 'Interoperability'.

Cybersecurity

resistance of an AI system to human-led, malicious external intrusions and manipulations that take place over general telecommunication networks and that primarily target the AI system itself and the datasets within its development rather than its inputs

Note 1. The 'primary' focus on the AI systems themselves and their databases serves to distinguish this test dimension from the 'Robustness' test dimension in the context of this definition set, insofar as Cybersecurity and Robustness are cited in the European AI Act in Art. 15 as two distinct requirements at the same level.

Note 2. The intersection with the 'Robustness' test dimension can be defined accordingly: it is found where intentional malicious intrusions are made in the inputs of an AI system.

Physical Security

resistance of an AI system to malicious external interventions and manipulations carried out by humans through the physical space

Robustness

ability of an AI system to maintain its regular and usual functioning and behaviour as best as possible within defined boundaries, even with atypical inputs and intended as well as primarily natural or random changes to them

Note 1. 'atypical inputs' can be understood to include unknown, adverse, disruptive or erroneous inputs or influences.

Note 2. The 'primary' focus on natural, random and therefore non-malicious changes serves to distinguish this test dimension from the 'Cybersecurity' test dimension in the context of this definition set, insofar as Cybersecurity and Robustness are cited in the European AI Act in Art. 15 as two distinct requirements at the same level. See also the notes on 'Cybersecurity'.

Explainability

property of an AI system with regard to the principle intelligibility and comprehensibility of functionality, behaviour and outputs, primarily for human specialists with a technical background in AI technologies, computer science, mathematics or comparable fields

Traceability

property of an AI system, with regard to the detectability of the consecutive states of the AI system during input processing and the outputs of the AI system

Transparency

property of an AI system with regard to the basic comprehensibility and accessibility of its internal components, parameters and functions, with regard to the entire life cycle as well as all components, tests and decisions that have been incorporated into its development and training

Note 1. The broad focus of this dimension reflects the respective tendency in ISO/IEC 22989:2002 (5.15.8).

Predictability

feasibility of reliable anticipation of future behaviour and functioning of an AI system, based on externally observable, rule-based consistency in the behaviour and output of the system

Note 1. In accordance with ISO/IEC 22989:2022 (5.15.7), 'Predictability' describes an epistemic relationship to the AI system that can support a relationship of trust without necessarily requiring an in-depth understanding of the internal functioning of the AI system.

Reproducibility

property of an AI system with regard to the generation of the same – or in specific cases: similar – outputs and behaviours shown, given the same or similar inputs and initial conditions

Note 1. In specific cases, namely when AI systems contain components with randomization elements, one can only speak of reproducibility – if at all – on the basis of similarities in the output. In other cases, equality must be assumed.

Observability

property of an AI system with regard to the gradual possibility for a human individual to monitor the behaviour or functioning of an AI system during operation

Interpretability

Understandability and comprehensibility of the functionality, behaviour and outputs of an AI system for people without a specific technical background in AI technologies, computer science, mathematics or comparable fields

Note 1. The choice of the term 'Interpretability' is based on Art. 13 (1) AI Act; notwithstanding of the tendency that 'Explainability' and 'Interpretability' are sometimes used as synonyms in certain discourses.

Privacy

inaccessibility of certain characteristics and areas of a human individual as such, including thoughts, beliefs, predispositions and orientations; as well as of certain physical spaces

Personal Data Protection

inaccessibility of certain characteristics of a human individual documented as information, including the control and influence of an indivudal over what information may be collected, stored and processed and who may disclose this information

Fundamental Rights

comprehension of all inalienable rights of a human individual, either on the basis of natural law or on the basis of positive law

Author's Personal Rights ("Urheberpersönlichkeitsrecht")

non-alienable personal legal position of the author or creator in his or her work, provided that the work has a sufficient degree of complexity

Note 1 The 'sufficient degree of complexity' is aimed at the concept of 'threshold of originality'.

Personalization

the degree to which the output or behaviour of an AI system is specifically adapted or responsive to a human individual and its behaviours, preferences, predispositions or characteristics

Nudging

unconscious, directed, suggestive effect of the output or behaviour of an AI system on an individual and its decisions or actions, where the effect is determined autonomously – in the sense of autonomy of ends – by a third party

Framing

property of the embedding of an AI system, in particular the user interface, with regard to the indication of the interactive or procedural involvement of an AI system

Controllability

property of an AI system, with regard to the gradual feasibility for a human individual to determine – be it in the sense of an autonomy of means or of ends – the behaviour or functioning of an AI system, in principle as well as during ongoing operation, and, if necessary, to stop it in an orderly manner

Usability

property of an AI system with regard to the quality of interaction and operation by a user, especially with regard to barrier-free usage

Non-Discrimination

property of an open process carried out by an AI system (consisting of: initial conditions, execution

and result), if in the course of this process several human individuals are treated in comparison to each other and/or act with each other, or if single individuals are treated, and in both cases this process is legally free of discrimination; whereby discrimination is understood as a less favourable treatment of a human individual on the basis of a legally protected property.

Note 1. Protected characteristics are defined in relevant legal texts; the German General Act on Equal Treatment e.g. defines a number of protected characteristics (race, ethnic origin, gender, religion or belief, disability, age, sexual identity; see para. 1 AGG).

Bias

1. (for datasets) a gradually distinct, directed deviation of a dataset compared to a reference dataset with regard to a specific aspect, thus a distorted distribution compared to a reference distribution

2. (for an actor's behaviour) a gradually distinct, directed and unconscious deviation in the behaviour (actions, judgements) of a an actor, which manifests itself in the accumulation as a deviation from a reference for this behaviour with regard to a specific aspect

Note 1. Provided that the second meaning contains a statistical element, the second meaning can be represented in the form of the first meaning – namely via the statistical recording of behaviour. The term 'systematic bias' is sometimes used for statistically detectable distorted behaviour.

Note 2. The use of the term 'bias' for AI parameters must be distinguished from the above-mentioned definitions of the term; in this case, 'bias' means a trainable parameter of an AI model that is included in the model as a constant.

Fairness

property of an open process carried out by an AI system (consisting of: initial conditions, execution and result), if in the course of this process several human individuals are treated in comparison to each other and/or act with each other, or if single individuals are treated, and in both cases this process corresponds to the non-legally established ideas and perceptions of justice (as well as of adjacent concepts) of groups or individuals to be named

Secrecy

property of an AI system with regard to the specific inaccessibility and non-disclosure of relevant corporate or government secrets

Factuality

property of the inputs and, in particular, the outputs of an AI system, with regard to the

correspondence of the respective data and information with the world-side facts formulated therein, especially where this correspondence is relevant to social processes and dynamics

Note 1. The present definition refers primarily to Large Language Models and specifically to the problem of hallucination (and subsequently in particular the relevance of this problem for social processes); in its generality, however, the definition is in principle also applicable to other AI technologies.

Marking

the overt or covert indication of the fact that the output of an AI system was generated entirely or primarily by AI

Copyright

rights of use and exploitation of a work, which are not necessarily held by the author of the work

Accountability

property of a human or legal person, with regard to the clearest possibility of attribution and resulting assumption of obligations for the effects of an AI system, in particular if third parties suffer damage or disadvantageous treatment as a result of an AI system

Reversibility

property of the embedding of an AI system, with regard to the possibility of restoring a state before the influence of the AI system on this embedding or the neighbouring environment, within a practicable period of time after this influence

Representativeness

property of a dataset with identical or sufficiently similar existence of a relational distribution in this dataset compared to a reference dataset with regard to a specific aspect

Accessibility

property of an AI system (or the underlying AI technology) with regard to the fundamental and financially practicable availability of this system or technology for human individuals from a global perspective

Workforce Exploitation

form of use of human labour in the life cycle of an AI system, including the hardware and infrastructure used for its purpose, that does not comply with the basic requirements for a humane work process

Perspectivity

the constitution of an AI system over its entire life cycle, with regard to the design of the influence of the system on humanity and human coexistence, in such a way that this life cycle – ceteris paribus – is also possible at future points in time under at least equivalent initial and general conditions

Resources

natural resources that are used or consumed within the life cycle of an AI system with regard to the underlying hardware and infrastructure

Energy

consumables used within the life cycle of an AI system for its processes or in the production of underlying hardware and infrastructure

Sustainability

constitution of an AI system over its entire life cycle, with regard to the design of the influence of the system on its natural environment, including the human habitat, in such a way that this life cycle – ceteris paribus – is also possible at future points in time under at least equivalent initial and general conditions

II. Life Cycle Steps

For the sake of readability, the following brief descriptions only refer to *one* AI model in the singular. It is nevertheless conceivable that an AI system could also contain several AI models.

Inception

Mission Statement

formulation of the goals and ends that the AI system should fulfill

KPIs Specification

concrete formulation of the central criteria through which the achievement of the purpose of the system can be recognized and measured

Requirements Engineering

formulation of the properties that the AI system must fulfill with regard to (further) expectations and specifications – for example, from the customer or with regard to legal requirements

Stakeholder Analysis

identification of all relevant individuals and interest groups involved in or affected by the development and operation of the AI system

Design & Analysis

Analysis & Approach

detailed analysis of the AI system's task and superordinate specification of the way in which this task is to be addressed

General Research

investigation of existing research approaches and solutions with regard to the formulated problem and the envisaged solutions

Model Research & Selection

investigation for possible model architectures, basic AI technologies and, if applicable, already pre-trained models; as well as the process of (provisionally) selecting a model (or a model architecture or an AI technology)

Data Concept

development of the basic concept for the data which which the AI system is supposed to draw and to build on during development and operation; including their specification

System Design

conception of the overall architecture of the AI system, in particular with regard to the constellation of the individual AI and non-AI components

Embedding Design

conception of the digital and, if applicable, analog environment into which the planned AI system is to be integrated

User Interface Design

conception of the interfaces for the interaction of end users and controlling persons with the AI system

Development

Data Collection & Selection

collection and selection of all necessary data according to the developed data concept through independent collection or by adopting existing data sets

Data Preprocessing & Cleaning

preparation of the entries in the data sets for further processing (e.g. by standardising formats, data types, etc.) and elimination of incorrect and otherwise unusable entries in the data sets

Data Labelling

process of assigning target variables (values, labels, etc.) to individual entries in the data sets

Data Augmentation

expansion of existing data sets through further data collection, enrichment with synthetic data and completion of incomplete entries

Hyperparameter Selection

(preliminary) determination of the aspects of the AI model that cannot be trained directly

Programming

implementation of the AI system in program code, with regard to the hard-coded parts of the system

(Model) Training

successive optimisation process of the changeable parameters of the AI model based on the training data and with regard to a defined optimisation target

Hyperparameter Tuning

improvement of aspects of the AI model that cannot be trained directly

Verification & Validation

Test Data Preparation

provision of the test datasets, either by performing the data processing steps separately (Data

Collection & Selection, Data Preprocessing & Cleaning, Data Labeling, Data Augmentation) or by suitable separation of the test data from the data pool *before* the training phase

Model Evaluation

evaluation of the selected model (and, if necessary, adjustment or return to previous phases)

System Verification

test process of the AI system for its general and AI-specific function using specific test procedures

Deployment

Virtual Deployment

inclusion of the AI system in a possibly broader target system, including possible interfaces for interaction with other virtual entities

Physical Deployment

inclusion of the AI system in the target hardware (possibly via the overall software system), including possible interfaces for physical interaction with the outside world

Model Deployment

depending on the specification, separate installation of the AI model on separate target hardware if necessary

Documentation & Manual

(final) written documentation of the development process, the system structure, the training and test procedures and the creation of written instructions for the use of the AI system

Note 1. In practice, these processes usually run parallel to the previous phases; however, deployment and delivery are the time when their results must actually be available.

Operation

Activity

procedural fact as such that an AI system is actually in operation and an AI model is being executed

Input Operation

processing the inputs for the AI system during operation

Output

statements and outputs of the AI system in the form of electronically coded data during operation

Behaviour

physical manifestations and effects of the overall system during operation

Interoperation

interaction of the AI system specifically with other AI components and AI systems during operation

Monitoring

Tracking

usually automated measurement in real time of values and parameters of the AI system during its operation and of the datasets used in the process

(Re-)Evaluation & Updating

repeated evaluation of the measured values and parameters of the AI system during operation (including automated test processes) as well as modifications to the AI system

Data Monitoring & Retraining

repeated evaluation of the datasets used in operation and, if necessary, modification of these datasets and further optimisation of the AI model (see Model Training)

Logging

usually automated storage of various information, values and parameters over time during the operation of the AI system

Incidence Detection

identification of undesirable and potentially problematic behaviour of the AI system or the resulting situations during operation

Incidence Mitigation

dealing with undesirable and potentially problematic behavior of the AI system or the resulting situations during operation

Incidence Reporting

notification of the fact that an undesirable or problematic behavior of the AI system has occurred to the appropriate parties

Retirement

Disposal

deletion of no longer required data, information and other resources from the entire life cycle as well as of the implemented AI system and the AI model

Archiving

long-term preservation of data, information and resources from the entire life cycle that are still required

TÜV AI.LAB

… throughout its entire life cycle.

Phase	Step	Consciousness	Spontaneity	Autonomy	Automation	Interoperability	Digital Operability	Physical Operability	Reliability	Performance	Safety
AI that is …		… safe, …									
Section		Regular Operation (AI System → Outside)									
Dimension		level of consciousness of an AI system based on inner mental states in a first person perspective or an analog basis	ability of an AI system to initiate a causal chain or activity in a targeted manner and according to its own criteria without a preceding or external impulse	ability of an AI system to independently determine its functionality and behaviour with regard to goals and ends ('autonomy of ends'), including changes to its own operational design domain, as well as the means and ways to achieve them ('autonomy of means') according to its own criteria	ability of an AI system to carry out independently multi-part, sufficiently complex processes in sequence, without further external support, in particular from human individuals	ability of an AI system to specifically influence other AI systems in various ways, both digitally and physically	ability of an AI system to realise its goals and ends and/or its ways and means in the digital space	ability of an AI system to realise its goals and purposes and/or its ways and means in the real world; this includes the ability to move in physical space	ability of an AI system to maintain its regular functionality and behaviour over time as consistently as possible and without failures in the event of malfunctions caused by internal components	measurable ability of an AI system to achieve the goals and ends which are given to it – heteronomously or autonomously – in the sense of a task	harmlessness of an AI system which is considered as sufficient for specified protection goals in the intended, functional operation
Inception	Mission Statement									DL Art. 11 (1) + An. IV (1 a), (2 b)	DL Art. 9 (1–2), (4–10)
	KPIs Specification									DL Art. 15 (1)	
	Requirements Engineering										
	Stakeholder Analysis										
Design & Concept	Analysis & Approach									DL Art. 11 (1) + An. IV (1), (2 a–b); Art. 15 (1)	
	General Research										
	Model Research & Selection										
	Data Concept										ML | DL Art. 10 (2 a)
	System Design								ML | DL | TL Art. 15 (4)	DL Art. 11 (1) + An. IV (2 b)	DL Art. 11 (1) + An. IV (2 b)
	Embedding Design										
	User Interface Design										
Development	Data Collection & Selection									DL Art. 11 (1) + An. IV (2 d)	ML | DL Art. 10 (2, f)
	Data Pre-processing & Cleaning										
	Data Labelling										
	Data Augmentation										
	Hyperparameter Selection									DL | TL Art. 15 (1)	
	Programming										
	(Model) Training										ML | DL Art. 10 (2, f)
	Hyperparameter Tuning										
Verification & Validation	Test Data Preparation									DL Art. 11 (1) + An. IV (2 g)	ML | DL Art. 10 (2, f)
	Model Evaluation										
	System Verification										
Deployment	Virtual Deployment					DL Art. 11 (1) + An. IV (1 b)	DL Art. 11 (1) + An. IV (1 b–e)	DL Art. 11 (1) + An. IV (1 b–e)			
	Physical Deployment										
	Model Deployment										
	Documentation & Manual					DL Art. 11 (1) + An. IV (1 b)	DL Art. 11 (1) + An. IV (1 b–e)	DL Art. 11 (1) + An. IV (1 b–e)		DL Art. 11 (1) + An. IV (2 a–c), (2 f), (3–4), (9); Art. 13 (3 b–e); Art. 15 (3); Art. 18 (1–2)	DL Art. 11 (1) + An. IV (2 a–c), (3), (5), Art. 13 (3 b); Art.18 (1–2)
Operation	Activity								ML | DL | TL Art. 15 (4)	DL | TL Art. 15 (1)	
	Input Operation										
	Output										
	Behaviour										
	Interoperation					DL | TL Art. 72 (2)					
Monitoring	Tracking									ML | DL | TL Art. 72 (2)	DL | TL Art. 72 (2)
	(Re-)Evaluation & Updating										ML Art. 73 (6) Art. 82 (1–2)
	Data Monitoring & Retraining										
	Logging										
	Incidence Detection										
	Incidence Mitigation										
	Incidence Reporting										ML Art. 73 (1–5)
Retirement	Disposal										
	Archiving									DL Art. 74 (12)	

Phase	Step	Unbreachability	Cybersecurity	Physical Security	Robustness	Explainability	Traceability	Transparency	Predictability	Reproducibility	Observability	Interpretability
AI that is …		… secure, …				… understandable, …						
Section		Disruption (AI System ← Outside)				Epistemology (AI System ← Individual)						
Dimension		resistance of an AI system to malicious external intrusions and manipulations, conducted primarily by AI systems in a means- or even ends-autonomous manner	resistance of an AI system to human-led, malicious external intrusions and manipulations that take place over general telecommunication networks and that primarily target the AI system itself and the datasets within its development rather than its inputs	resistance of an AI system to malicious external interventions and manipulations carried out by humans through the physical space	ability of an AI system to maintain its regular and usual functioning and behaviour as best as possible within defined boundaries, even with atypical inputs and intended as well as primarily natural or random changes to them	property of an AI system with regard to the principle intelligibility and comprehensibility of functionality, behaviour and outputs, primarily for human specialists with a technical background in AI technologies, computer science, mathematics or comparable fields	property of an AI system, with regard to the detectability of the consecutive states of the AI system during input processing and the outputs of the AI system	property of an AI system with regard to the basic comprehensibility and accessibility of its internal components, parameters and functions, with regard to the entire life cycle as well as all components, tests and decisions that have been incorporated into its development and training	feasibility of reliable anticipation of future behaviour and functioning of an AI system, based on externally observable, rule-based consistency in the behaviour and output of the system	property of an AI system with regard to the generation of the same – or in specific cases: similar – outputs and behaviours shown, given the same or similar inputs and initial conditions	property of an AI system with regard to the gradual possibility for a human individual to monitor the behaviour or functioning of an AI system during operation	understandability and comprehensibility of the functionality, behaviour and outputs of an AI system for people without a specific technical background in AI technologies, computer science, mathematics or comparable fields
Inception	Mission Statement		DL Art. 15 (1)		DL Art. 15 (1)			DL Art. 13 (1)				
	KPIs Specification										DL Art. 14 (1–4)	DL Art. 13 (1)
	Requirements Engineering											
	Stakeholder Analysis											
Design & Concept	Analysis & Approach		DL Art. 11 (1) + An. IV (1), (2 a–b); Art. 15 (1), (5)		DL Art. 11 (1) + An. IV (1), (2 a–b); Art. 15 (1), (4)			DL Art. 11 (1) + An. IV (1), (2 a–b); Art. 13 (1)			DL Art. 11 (1) + An. IV (1), (2 a–b); Art. 14 (1–4)	DL Art. 11 (1) + An. IV (1), (2 a–b); Art. 13 (1); Art. 14 (4)
	General Research											
	Model Research & Selection											
	Data Concept											
	System Design		DL Art. 11 (1) + An. IV (2 b), (2 f)		DL Art. 11 (1) + An. IV (2 b)		DL Art. 12 (1–2)	DL Art. 13 (1)			DL Art. 14 (1–4)	DL Art. 13 (1); Art. 14 (4)
	Embedding Design											
	User Interface Design											
Development	Data Collection & Selection		DL Art. 11 (1) + An. IV (2 d)		DL Art. 11 (1) + An. IV (2 d)			DL Art. 13 (1)				
	Data Pre-processing & Cleaning											
	Data Labelling											
	Data Augmentation											
	Hyperparameter Selection		DL | TL Art. 15 (1)		DL | TL Art. 15 (1), (4)							
	Programming		DL Art. 11 (1) + An. IV (2 f)								DL | TL Art. 14 (1–3)	DL | TL Art. 13 (1); Art. 14 (4)
	(Model) Training		ML | DL | TL Art. 15 (5)									
	Hyperparameter Tuning											
Verification & Validation	Test Data Preparation				DL Art. 11 (1) + An. IV (2 g)							
	Model Evaluation											
	System Verification											
Deployment	Virtual Deployment											
	Physical Deployment											
	Model Deployment											
	Documentation & Manual		DL Art. 11 (1) + An. IV (2 a–c), Art. 13 (3 b); Art. 18 (1–2)		DL Art. 11 (1) + An. IV (2 a–c), (3); Art. 13 (3 b); Art. 18 (1–2)		DL Art. 11 (1) + An. IV (2 a–b); Art. 13 (3 f); Art. 18 (1–2); Art. 72 (3)	DL Art. 11 (1) + An. IV (2 a–b); Art. 13 (2); Art. 18 (1–2)			DL Art. 11 (1) + An. IV (3)	DL Art. 11 (1) + An. IV (2 a–b); Art. 13 (2–3); Art. 18 (1–2)
Operation	Activity		DL | TL Art. 15 (1)		DL | TL Art. 15 (1), (4)			DL Art. 13 (1)				
	Input Operation		DL | TL Art. 15 (5)				DL Art. 12 (3 c)				ML Art. 26 (4)	
	Output											
	Behaviour											
	Interoperation											
Monitoring	Tracking				ML Art. 15 (4)		DL | TL Art. 72 (2)				ML Art. 14 (4); Art. 26 (1–2), (5)	ML Art. 14 (4)
	(Re-)Evaluation & Updating											
	Data Monitoring & Retraining											
	Logging						ML | DL | TL Art. 12 (2–3)					
	Incidence Detection											
	Incidence Mitigation											
	Incidence Reporting											
Retirement	Disposal											
	Archiving		DL Art. 74 (12)		DL Art. 74 (12)		DL Art. 16 (e); Art. 19; Art. 26 (6); Art. 74 (12)	DL Art. 74 (12)				

Phase	Step	Privacy	Personal Data Protection	Fundamental Rights	Author's Personal Rights	Personalization	Nudging	Framing	Controllability	Usability
AI that is …		… respectful and controllable, …								
Section		Individual Person (AI System ↔ Individual)								
Dimension		inaccessibility of certain characteristics and areas of a human individual as such, including thoughts, beliefs, predispositions and orientations; as well as of certain physical spaces	inaccessibility of certain characteristics of a human individual documented as information, including the control and influence of an indivudal over what information may be collected, stored and processed and who may disclose this information	comprehension of all inalienable rights of a human individual, either on the basis of natural law or on the basis of positive law	non-alienable personal legal position of the author or creator in his or her work, provided that the work has a sufficient degree of complexity	the degree to which the output or behaviour of an AI system is specifically adapted or responsive to a human individual and its behaviours, preferences, predispositions or characteristics	unconscious, directed, suggestive effect of the output or behaviour of an AI system on an individual and its decisions or actions, where the effect is determined autonomously – in the sense of autonomy of ends – by a third party	property of the embedding of an AI system, in particular the user interface, with regard to the indication of the interactive or procedural involvement of an AI system	property of an AI system, with regard to the gradual feasibility for a human individual to determine – be it in the sense of an autonomy of means or of ends – the behaviour or functioning of an AI system, in principle as well as during ongoing operation, and, if necessary, to stop it in an orderly manner	property of an AI system with regard to the quality of interaction and operation by a user, especially with regard to barrier-free usage
Inception	Mission Statement									
	KPIs Specification								DL Art. 14 (1–4)	
	Requirements Engineering									
	Stakeholder Analysis									
Design & Concept	Analysis & Approach								DL Art. 11 (1) + An. IV (1), (2 a–b); Art. 14 (1–4)	
	General Research									
	Model Research & Selection									
	Data Concept		ML | DL Art. 10 (2 a)	ML | DL Art. 10 (2 a)						
	System Design							TL Art. 50 (2)	DL Art. 14 (1–4)	
	Embedding Design									
	User Interface Design							TL Art. 50 (1), (3)	DL Art. 11 (1)+ An.IV (1 g)	TL Art. 16 (l)
Development	Data Collection & Selection		ML | DL Art. 10 (2), (5)	ML | DL Art. 10 (2, f)						
	Data Pre-processing & Cleaning									
	Data Labelling									
	Data Augmentation									
	Hyperparameter Selection									
	Programming								DL Art. 14 (1–4)	
	(Model) Training		ML | DL Art. 10 (2), (5)	ML | DL Art. 10 (2)						
	Hyperparameter Tuning									
Verification & Validation	Test Data Preparation		ML | DL Art. 10 (2), (5)	ML | DL Art. 10 (2, f)						
	Model Evaluation									
	System Verification									
Deployment	Virtual Deployment									
	Physical Deployment									
	Model Deployment									
	Documentation & Manual		DL Art. 47 (2)	DL Art. 27 (1); Art. 11 (1) + An. IV (3)					DL Art. 11 (1) + An. IV (2 a–b, e), (3); Art. 13 (2 b–c)	
Operation	Activity							TL Art. 50 (1), (3)		TL Art. 50 (5)
	Input Operation		TL Art. 50 (3)							
	Output									
	Behaviour									
	Interoperation									
Monitoring	Tracking								ML Art. 14 (4); Art. 26 (1–2), (5)	
	(Re-)Evaluation & Updating									
	Data Monitoring & Retraining									
	Logging									
	Incidence Detection									
	Incidence Mitigation									
	Incidence Reporting			ML Art. 73 (9–10)						
Retirement	Disposal		ML Art. 10 (2), (5, e–f), (6)							
	Archiving									

Phase	Step	Non-Discrimination	Bias	Fairness	Secrecy	Factuality	Marking	Copyright	Accountability	Reversibility	Representativeness
AI that is …		… just, …			… for the benefit of society, …						
Section		Several Individuals (AI System → Individuals)			Society (AI System → All Individuals of a Society)						
Dimension		property of an open process carried out by an AI system (consisting of: initial conditions, execution and result), if in the course of this process several human individuals are treated in comparison to each other and/or act with each other, or if single individuals are treated, and in both cases this process is legally free of discrimination; whereby discrimination is understood as a less favourable treatment of a human individual on the basis of a legally protected property	1. (for datasets) a gradually distinct, directed deviation of a dataset compared to a reference dataset with regard to a specific aspect, thus a distorted distribution compared to a reference distribution 2. (for an actor's behaviour) a gradually distinct, directed and unconscious deviation in the behaviour (actions, judgements) of an actor, which manifests itself in the accumulation as a deviation from a reference for this behaviour with regard to a specific aspect	property of an open process carried out by an AI system (consisting of: initial conditions, execution and result), if in the course of this process several human individuals are treated in comparison to each other and/or act with each other, or if single individuals are treated, and in both cases this process corresponds to the non-legally established ideas and perceptions of justice (as well as of adjacent concepts) of groups or individuals to be named	property of an AI system with regard to the specific inaccessibility and non-disclosure of relevant corporate or government secrets	property of the inputs and, in particular, the outputs of an AI system, with regard to the correspondence of the respective data and information with the world-side facts formulated therein, especially where this correspondence is relevant to social processes and dynamics	the overt or covert indication of the fact that the output of an AI system was generated entirely or primarily by AI	rights of use and exploitation of a work, which are not necessarily held by the author of the work	property of a human or legal person, with regard to the clearest possibility of attribution and resulting assumption of obligations for the effects of an AI system, in particular if third parties suffer damage or disadvantageous treatment as a result of an AI system	property of the embedding of an AI system, with regard to the possibility of restoring a state before the influence of the AI system on this embedding or the neighbouring environment, within a practicable period of time after this influence	property of a dataset with identical or sufficiently similar existence of a relational distribution in this dataset compared to a reference dataset with regard to a specific aspect
Inception	Mission Statement										
	KPIs Specification										
	Requirements Engineering										
	Stakeholder Analysis										
Design & Concept	Analysis & Approach										
	General Research										
	Model Research & Selection										
	Data Concept	ML | DL Art. 10 (2, a)	ML | DL Art. 10 (2, a)				ML | DL Art. 10 (2 h)				ML | DL Art. 10 (2 a)
	System Design		DL Art. 15 (4)				TL Art. 50 (2)				
	Embedding Design										
	User Interface Design										
Development	Data Collection & Selection	ML | DL Art. 10 (2, f); Art. 11 (1)+ An. IV (2 d)	ML | DL Art. 10 (2, f–g), (5); Art. 11 (1)+ An. IV (2 d)				ML | DL Art. 10 (2 h)				ML | DL Art. 10 (3–4)
	Data Pre-processing & Cleaning	ML | DL Art. 10 (2 h)	ML | DL Art. 10 (2 h)								
	Data Labelling										
	Data Augmentation										
	Hyperparameter Selection		ML | DL | TL Art. 15 (4)								
	Programming										
	(Model) Training	ML | DL Art. 10 (2, f)	ML | DL Art. 10 (2, f–g), (5)								ML | DL Art. 10 (3–4)
	Hyperparameter Tuning										
Verification & Validation	Test Data Preparation	ML | DL | TL Art. 10 (2, f); Art. 11 (1) + An. IV (2 g), (3)	ML | DL | TL Art. 10 (2, f–g), (5); Art. 11 (1) + An. IV (2 g), (3)				ML | DL Art. 10 (2 h)				ML | DL Art. 10 (3–4)
	Model Evaluation										
	System Verification										
Deployment	Virtual Deployment										
	Physical Deployment										
	Model Deployment										
	Documentation & Manual	DL Art. 11 (1) + An. IV (2 a–c); Art. 13 (2)	DL Art. 11 (1) + An. IV (2 a–c); Art. 13 (2)						DL Art. 11 (1) + An. IV (8); Art. 47 (1–2)		
Operation	Activity						ML Art. 26 (7), (11)				
	Input Operation		ML | DL | TL Art. 15 (4)								
	Output						TL Art. 50 (2), (4)				
	Behaviour										
	Interoperation										
Monitoring	Tracking										
	(Re-)Evaluation & Updating										
	Data Monitoring & Retraining										
	Logging										
	Incidence Detection										
	Incidence Mitigation										
	Incidence Reporting										
Retirement	Disposal										
	Archiving	DL Art. 74 (12)	DL Art. 74 (12)						DL Art. 47 (1)		DL Art. 74 (12)

Phase	Step	Accessibility	Workforce Exploitation	Perspectivity	Resources	Energy	Sustainability
AI that is …		… humanity …			… and our planet …		
Section		Global Community (AI Sys. → Humanity)			Global Ecosystem (AI System → Planet)		
Dimension		property of an AI system (or the underlying AI technology) with regard to the fundamental and financially practicable availability of this system or technology for human individuals from a global perspective	form of use of human labour in the life cycle of an AI system, including the hardware and infrastructure used for its purpose, that does not comply with the basic requirements for a humane work process	the constitution of an AI system over its entire life cycle, with regard to the design of the influence of the system on humanity and human coexistence, in such a way that this life cycle – ceteris paribus – is also possible at future points in time under at least equivalent initial and general conditions	natural resources that are used or consumed within the life cycle of an AI system with regard to the underlying hardware and infrastructure	consumables used within the life cycle of an AI system for its processes or in the production of underlying hardware and infrastructure	constitution of an AI system over its entire life cycle, with regard to the design of the influence of the system on its natural environment, including the human habitat, in such a way that this life cycle – ceteris paribus – is also possible at future points in time under at least equivalent initial and general conditions
Inception	Mission Statement						
	KPIs Specification						
	Requirements Engineering						
	Stakeholder Analysis						
Design & Concept	Analysis & Approach						
	General Research						
	Model Research & Selection						
	Data Concept						
	System Design						
	Embedding Design						
	User Interface Design						
Development	Data Collection & Selection						
	Data Pre-processing & Cleaning						
	Data Labelling						
	Data Augmentation						
	Hyperparameter Selection						
	Programming						
	(Model) Training						
	Hyperparameter Tuning						
Verification & Validation	Test Data Preparation						
	Model Evaluation						
	System Verification						
Deployment	Virtual Deployment						
	Physical Deployment						
	Model Deployment						
	Documentation & Manual						
Operation	Activity						
	Input Operation						
	Output						
	Behaviour						
	Interoperation						
Monitoring	Tracking						
	(Re-)Evaluation & Updating						
	Data Monitoring & Retraining						
	Logging						
	Incidence Detection						
	Incidence Mitigation						
	Incidence Reporting						
Retirement	Disposal						
	Archiving						

High-Risk-Requirements
Transparency-Requirements
… with regard to AI systems

TL = Test Layer
DL = Documentation Layer
ML = Management Layer

The QMS (Art. 17) is part of the ML in the respective areas