IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 7 December 2023, accepted 29 January 2024, date of publication 8 February 2024, date of current version 15 February 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3364387

RESEARCH ARTICLE

Trustworthiness Assurance Assessment for High-Risk AI-Based Systems

GEORG STETTINGER[1], PATRICK WEISSENSTEINER[2], AND SIDDARTHA KHASTGIR[3], (Member, IEEE)

[1]Infineon Technologies AG, 85579 Neubiberg, Germany
[2]Virtual Vehicle Research GmbH, 8010 Graz, Austria
[3]International Manufacturing Centre, The University of Warwick (WMG), CV4 7AL Coventry, U.K.

Corresponding author: Georg Stettinger (georg.stettinger@infineon.com)

This work was supported by the European Union's Horizon Europe Research and Innovation Program under Grant 101076754—AIthena Project.

ABSTRACT This work proposes methodologies for ensuring the trustworthiness of high-risk artificial intelligence (AI) systems (AIS) to achieve compliance with the European Union's (EU) AI Act. High-risk classified AIS must fulfill seven requirements to be considered trustworthy and human-centric, and subsequently be considered for deployment. These requirements are equally important, mutually supportive, and should be implemented and evaluated throughout the AI lifecycle. The assurance of trustworthiness is influenced by ethical considerations, amongst others. Hence, the operational design domain (ODD) and behavior competency (BC) concepts from the automated driving domain are utilized in risk assessment strategies to quantify different types of residual risks. The methodology presented is guided by the consistent application of the ODD and its related BC concept throughout the entire AI lifecycle, focusing on the trustworthiness assurance framework and its associated process as the main pillars for AIS certification. The achievement of the overall objective of trustworthy and human-centric AIS is divided into seven interconnected sub-goals: the formulation of use restrictions, the trustworthiness assurance/argument itself, the identification of dysfunctional cases, the utilization of scenario databases and datasets, the application of metrics for evaluation, the implementation of the proposed concept across the AI lifecycle, and sufficient consideration of human factors. The role of standards in the assurance process is discussed, considering any existing gaps and areas for improvement. The work concludes with a summary of the developed approach, highlighting key takeaways and action points. Finally, a roadmap to ensure trustworthy and human-centric behavior of future AIS is outlined.

INDEX TERMS AI Act, AI life-cycle, assurance, behavior competencies, certification, dysfunctional cases, ethics, human-centric, KPIs, metrics, operational design domain, residual risk, restrictions of use, risk assessment, standards, trustworthy.

I. INTRODUCTION

Artificial Intelligence (AI) technologies are anticipated to yield diverse economic and societal benefits across sectors such as the environment, health, public sector, finance, mobility, home affairs, and agriculture [1]. They prove especially valuable in enhancing prediction accuracy, optimizing operations and resource allocation, and tailoring services to individual needs. Nevertheless, concerns arise regarding the impact of Artificial Intelligence Systems (AIS) on fundamental rights safeguarded by the EU Charter of Fundamental Rights. Additionally, safety and security risks for users emerge when AI technologies are integrated into products and services. In particular, AIS may compromise fundamental rights such as the right to non-discrimination, freedom of expression, human dignity, protection of personal data, and privacy [2].

Furthermore, the rapid evolution of general-purpose artificial intelligence (GPAI) technologies [3], exemplified by ChatGPT [4], [5], is reshaping the landscape of AI system

The associate editor coordinating the review of this manuscript and approving it for publication was Amjad Mehmood.

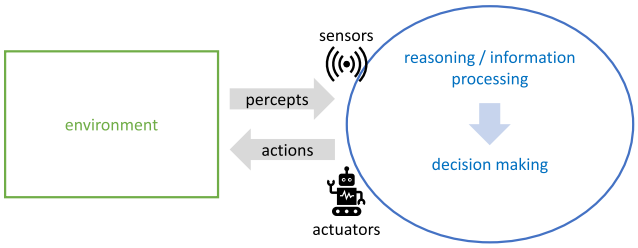

FIGURE 1. A schematic depiction of an AI system [10].

development and deployment [6]. While these technologies promise substantial benefits in the years ahead, fostering innovation across various sectors, their disruptive nature raises policy concerns related to privacy, intellectual property rights, liability, accountability, and the potential spread of disinformation and misinformation [7]. EU lawmakers must navigate a careful balance between promoting the use of these technologies and ensuring the implementation of appropriate safeguards [8].

The majority of these systems and technologies are expected to operate in an open context environment. According to [9], an open context introduces complexity and unpredictability to the deployment environment, influencing the validation of these systems and technologies. Therefore, there is a need for innovative concepts to describe the open context, including its boundaries, and articulate the capabilities of these systems within the deployment environment.

A. DEFINITION OF AI AND AI TAXONOMY

In June 2018, the European Commission established the AI High-Level Expert Group (AI HLEG) as an independent expert body. In 2019, one of their key outputs, "A Definition of AI: Main Capabilities and Disciplines" [10], aimed to formulate a unified definition of AI. The starting point was the definition of Artificial Intelligence (AI) proposed in the European Commission's Communication on AI [11], which states: *"Artificial intelligence (AI) refers to systems that display intelligent behaviour by analysing their environment and taking actions - with some degree of autonomy - to achieve specific goals. AI-based systems can be purely software-based, operating in the virtual world (e.g., voice assistants, image analysis software, search engines, speech and face recognition systems), or AI can be embedded in hardware devices (e.g., advanced robots, autonomous cars [12], drones, or Internet of Things applications)."*

Figure 1 illustrates the main features of this definition, emphasizing that any intelligent behavior of an AI system is rooted in perceiving and analyzing the environment before deciding and implementing corresponding actions. Moreover, this behavior can be realized as pure software operating in a virtual world or integrated into hardware devices.

In addition to the initial definition, two related classifications of AI systems have commonly been used over the last decade:

- A general (or strong) AI system is intended to perform most activities that humans can do.
- Narrow (or weak) AI systems, on the other hand, are designed to perform one or a few specific tasks.

As a culmination of their study, the AI HLEG updated the initial definition, incorporating additional details related to reasoning and decision-making in response to change requests from the scientific community. Specifically, the revised definition of AI is as follows:

"Artificial intelligence (AI) systems are software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal. AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their behaviour by analysing how the environment is affected by their previous actions. As a scientific discipline, AI includes several approaches and techniques, such as machine learning, machine reasoning, and robotics."

In addition to the AI HLEG deliverable, two additional reports on defining AI were published within the framework of AI Watch [13], the European Commission's knowledge service launched in December 2018 to monitor the development, uptake, and impact of Artificial Intelligence (AI) in Europe. The first report proposes an operational definition of AI to be adopted within the context of AI Watch. This operational definition comprises a concise taxonomy (AI domain and AI subdomain) and a list of keywords that characterize the core areas of AI research, along with addressing cross-cutting issues such as applications, ethical considerations, and philosophical aspects. The development of the operational definition is rooted in the definition of AI adopted by the AI HLEG. About a year later, AI Watch launched the second edition of the report [14]. Building on the first report from 2020, this edition incorporates several recent developments. Notably, the European Commission has proposed a regulatory framework for artificial intelligence (AI Act, see section I-B), which includes a legal definition of AI. In December 2023, the Council Presidency and the European Parliament negotiators reached a provisional agreement on the EU AI Act. According to the provisional agreement, the AI Act will apply two years after its entry into force, with some exceptions for specific provisions. Following the provisional agreement, work will continue at technical level in the coming weeks and months to finalise the details of the new regulation. Once this work has been completed, the Presidency will submit the compromise text to Member States' representatives for approval. The whole text will have to be confirmed by both institutions and undergo legal-linguistic revision before being formally adopted by the co-legislators [15].

Despite the multiple facets of AI and the consequent lack of a common definition, several commonalities are observed

in the analyzed definitions. The expression of these common aspects suggests that they can be considered as the main characteristics of AI:

- Perception of the environment, including the consideration of the real world complexity.
- Information processing: collecting and interpreting inputs (in the form of data).
- Decision making (including reasoning and learning): taking actions, performing tasks (including adaptation, reacting to changes in the environment) with a certain level of autonomy.
- Achievement of specific goals: considered as the ultimate purpose of AI systems.

Furthermore, the second edition of the report briefly presents alternative approaches to the study of AI. These approaches include classifying AI according to families of algorithms and the theoretical models behind them, cognitive abilities reproduced by AI, and functions performed by AI. AI applications can also be grouped according to other dimensions, such as the economic sector in which they are found or their business functions.

In November 2023, the Council of the Organisation for Economic Co-operation and Development (OECD) published a new definition of artificial intelligence [16]. This definition is slated to be included in the new EU AI rulebook and is likely to find a place in the upcoming EU AI regulation. As discussed earlier, the definition is a critical aspect as it delineates the scope of AI itself. Specifically, the new definition is as follows: *"An AI system is a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that [can] influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment."* The primary motivation for this new definition was to achieve international alignment of AI definitions, taking into account developments over the last five years, improve technical accuracy and clarity, and create a more 'future-proof' definition. One significant change is the removal of the requirement for goals to be defined by humans, allowing coverage of cases where the AI system can learn new goals. In summary, the multi-domain nature of the AI definition itself makes harmonization a challenging task. Each domain aims to have its specific aspects covered, leading to numerous discussions on how much detail is necessary and sufficient to include in the definition. The common aspects identified above underscore the strong connection with the domain of robotics and its underlying operating principle known as sense-plan-act, which also forms the basis of ADS.

B. THE AI ACT

1) INTRODUCTION

Given the rapid development of AI technologies, the regulation of AI has emerged as a crucial policy issue in the European Union (EU) in recent years [17]. Policymakers have committed to establishing a 'human-centric' approach to AI, ensuring that Europeans can derive benefits from new technologies developed in accordance with EU values and principles [18]. While the EU currently lacks a specific legal framework for AI, the EC's White Paper on Artificial Intelligence underscores the necessity for a regulatory and investment-oriented approach with the dual objectives of promoting AI adoption and addressing the risks associated with specific uses of this technology [1].

To achieve these goals, the EC initially embraced a soft-law approach by issuing its non-binding Ethical Guidelines for Trustworthy AI and providing policy and investment recommendations in 2019 [19]. However, in 2021, with the Communication on Fostering a European Approach to Artificial Intelligence, the Commission shifted to a legislative approach and advocated for the establishment of a new regulatory framework for artificial intelligence. As the existing legislation, designed to safeguard fundamental rights and ensure safety and consumer rights (including data protection and non-discrimination laws), appears insufficient to address the risks posed by AI technologies, the Commission proposes the adoption of harmonized rules governing the development, market placement, and use of AI systems. These new rules would complement and follow the logic of existing EU rules on safety products, and would be adopted alongside a new Machinery Regulation to adapt safety rules to a new generation of products, such as 3D printers.

The overarching goal of the proposed AI Act is to ensure the smooth operation of the internal market by establishing conditions for the development and utilization of reliable AI systems in the Union. The draft act establishes a harmonized legal framework for the development, market placement, and use of AI products and services. Additionally, the proposed AI Act aims to achieve several specific objectives:

- Ensure that AI systems placed on the EU market are safe and compliant with existing EU law.
- Provide legal certainty to facilitate investment and innovation in AI.
- Enhance governance and enforce EU law effectively concerning fundamental rights and safety requirements applicable to AI systems.
- Facilitate the development of a single market for legal, safe, and trustworthy AI applications and prevent market fragmentation.

The new AI framework would incorporate a technology-neutral definition of AI systems (AIS) and adopt a risk-based approach, stipulating varying requirements and obligations for the development, market placement, and use of AI systems in the EU. In practice, the proposal sets common mandatory requirements for the design and development of AI systems before they are placed on the market and harmonizes the process of ex-post controls. The proposed AI legislation would complement existing and future horizontal and sectoral EU safety legislation. The Commission suggests following the logic of the New Legislative Framework (NLF), which is the EU's approach to ensuring that a range of

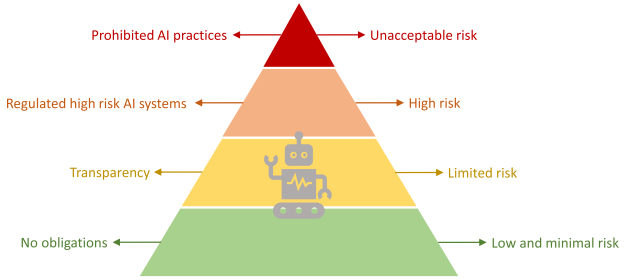

FIGURE 2. Pyramid of risks [1].

products complies with applicable legislation upon entering the EU market.

The new rules would primarily apply to providers of AIS established in the EU or in a third country that place AI systems on the EU market or put them into service in the EU, as well as to users of AIS established in the EU.

2) RISK-BASED APPROACH

The use of AI, with its specific characteristics (e.g., opacity, complexity, dependency on data, autonomous behavior), can adversely affect a number of fundamental rights and users' safety. To address those concerns, the draft AI Act follows a risk-based approach whereby legal intervention is tailored to the concrete level of risk [1]. To that end, the draft AI act distinguishes between AI systems posing:

- unacceptable risk,
- high risk,
- limited risk,
- and low or minimal risk.

Under this approach, AI applications would be regulated only as strictly necessary to address specific levels of risk.

Unacceptable risk: Prohibited AI practices. The proposed AI Act explicitly prohibits harmful AI practices that are considered to pose a clear threat to the safety, livelihoods, and rights of humans, because they pose an 'unacceptable risk'. Accordingly, they would be prohibited from being placed on the market, incorporated into services or used in the EU:

- AIS that deploy harmful manipulative 'subliminal techniques',
- AIS that exploit specific vulnerable groups (physical or mental disability),
- AIS used by public authorities, or on their behalf, for social scoring purposes,
- 'Real-time' remote biometric identification systems in publicly accessible spaces for law enforcement purposes, except in a limited number of cases.

High risk: Regulated high-risk AI systems. The proposed AI Act regulates 'high-risk' AI systems that have a negative impact on human safety or fundamental rights. The draft text distinguishes between two categories of high-risk AI systems.

- High-risk AIS used as a safety component of a product or as a product covered by EU health and safety harmonization legislation (e.g., toys, aviation, cars, medical devices, lifts).
- High-risk AIS used in eight specific areas identified in the Annex to the AI Act, which the Commission would be empowered to update as necessary by delegated act.

All these high-risk AI systems would be subject to a set of new rules, including

- **Requirement for an ex-ante conformity assessment:** Providers of high-risk AI systems would be required to register their systems in an EU-wide database managed by the Commission before placing them on the market or putting them into service. All AI products and services that are covered by existing product safety legislation will be covered by existing third-party conformity frameworks that already apply (e.g., for medical devices). Providers of AIS not currently covered by EU legislation would have to carry out their own conformity assessment (self-assessment) to demonstrate that they comply with the new requirements for high-risk AIS and can use the CE marking. Only high-risk AIS used for biometric identification would require conformity assessment by a 'notified body'.
- **Other requirements for high-risk AI systems:** Such systems would have to meet a number of requirements to ensure a trustworthy behavior, in particular with regard to
  - Human Agency and Oversight,
  - Technical Robustness and Safety,
  - Privacy and Data Governance,
  - Transparency,
  - Diversity, Non-discrimination and Fairness,
  - Social and Environmental Well-being,
  - and Accountability.

These seven key requirements represent the core elements of trustworthy AIS, which are detailed in the so-called Assessment List for Trustworthy AI [20]. This document was also produced by the AI HLEG. It is the third deliverable of the AI HLEG and follows the publication of the group's deliverable, the Ethics Guidelines for Trustworthy AI, which were published in 2019. Further details on Trustworthy AI are outlined in section II-A.

In this respect, suppliers, importers, distributors, and users of high-risk AIS would have to comply with a number of obligations. Suppliers established outside the Union will have to appoint an authorized representative in the EU to ensure conformity assessment, to set up a post-market surveillance system and to take corrective action where necessary. AIS that comply with expected new harmonized EU standards currently under development would benefit from a presumption of conformity with the requirements of the draft AI legislation.

Limited risk: Transparency obligations. A limited set of transparency obligations would apply to AIS presenting 'limited risk', such as systems that interact with humans (i.e., chatbots), emotion recognition systems, biometric categorization systems, and AIS that generate or manipulate image, audio or video content (i.e., deepfakes).

Low or minimal risk: No obligations. All other low or minimal risk AIS could be developed and used in the EU without additional legal obligations. However, the proposed AI act foresees the establishment of codes of conduct to encourage providers of non-high-risk AIS to voluntarily apply the mandatory requirements for high-risk AIS.

3) GOVERNANCE, ENFORCEMENT AND SANCTIONS

At the Union level, the proposal establishes the European Artificial Intelligence Board, composed of representatives from the Member States and the European Commission. This board aims to facilitate the harmonized implementation of the new rules and ensure cooperation between national supervisory authorities and the Commission. At the national level, Member States are required to designate one or more competent authorities, including a national supervisory authority, responsible for overseeing the application and implementation of the Regulation [1].

National market surveillance authorities are tasked with assessing operators' compliance with the obligations and requirements for high-risk Artificial Intelligence Systems (AIS). These authorities have access to confidential information, including the source code of the AIS, and are thus bound by confidentiality obligations. Additionally, they must take all necessary corrective measures to prohibit, restrict, withdraw, or recall AIS that do not comply with the requirements of the AI Act. Even compliant AIS that pose a risk to health, safety, fundamental rights, or other public interests must be addressed. In case of persistent non-compliance, concerned Member States must take appropriate measures to restrict, prohibit, recall, or withdraw the high-risk AIS from the market. Administrative fines, varying based on the seriousness of the infringement, are outlined to sanction non-compliance with the AI Act. Member States are obligated to establish rules on penalties, including administrative fines, and must take all necessary measures to ensure their proper and effective enforcement.

4) MEASURES TO SUPPORT INNOVATION

The Commission proposes that Member States or the European Data Protection Supervisor could set up a regulatory sandbox [21], i.e., a controlled environment that facilitates the development, testing, and validation of innovative AI (for a limited period of time) before it is put on the market [1]. The sandbox allows participants to use personal data to foster AI innovation, without prejudice to GDPR requirements. Other measures are specifically tailored to small providers and start-ups.

Reference [22] summarizes the results of a pre-regulatory sandbox with 9 European start-ups and SMEs. Participants in the sandbox were generally cautiously optimistic about the need for AI regulation [23]. However, the sandbox has highlighted a number of areas where improvements and further reflection are needed in order for participants to be able to assess the impact of the proposed AI legislation on their business operations.

II. MOTIVATION

The overarching objective is to develop Artificial Intelligence Systems (AIS) that seamlessly integrate trustworthiness and human-centricity, fostering an environment conducive to innovation. Central to this goal is the establishment of ethical principles serving as the core for Trustworthy AIS. To realize this vision, the Operational Design Domain (ODD) concept and its associated behavioral competencies play a pivotal role. This approach holds significant promise in guiding AIS towards a balanced integration of reliability, ethics, and innovation. A heightened awareness of potential risks, coupled with a meticulous assessment and subsequent mitigation, becomes a potent catalyst in the pursuit of a trustworthy and human-centered AIS. The strategic emphasis on risk management creates a proactive environment, nurturing AIS that instill confidence while addressing human needs. The culmination of this strategy is the amalgamation of residual risks and specific risk acceptance criteria, resulting in a landscape where AIS can be deployed with trustworthiness and confidence, upholding their commitment to both innovation and human welfare.

A. ETHICS AND TRUSTWORTHY AI

In 2019 the AI HLEG, published the Ethics Guidelines for Trustworthy Artificial Intelligence [10]. As a major outcome Trustworthy AIS were specified based on ethical principles. In that sense, trustworthy AI is built on three components that should be met throughout the system's lifecycle:

- it should be lawful, complying with all applicable laws and regulations,
- it should be ethical, ensuring compliance with ethical principles and values,
- and it should be robust, both technically and socially, as even with good intentions, AI systems can cause unintended harm.

Each component is necessary, but not sufficient, to achieve Trustworthy AI. Ideally, all three components will work in harmony and overlap. If, in practice, tensions arise between these components, society should seek to reconcile them. These Guidelines provide a framework for achieving Trustworthy AI, see Figure 3.

1) FOUNDATIONS OF TRUSTWORTHY AI

The foundations of Trustworthy AI are based on fundamental rights as moral and legal entitlements [24], [25] and are reflected in four ethical principles, namely

- Respect for human autonomy,
- Prevention of harm,
- Fairness,
- and Explicability,

which should be respected to ensure ethical and robust AI. Tensions between the above principles may arise and there

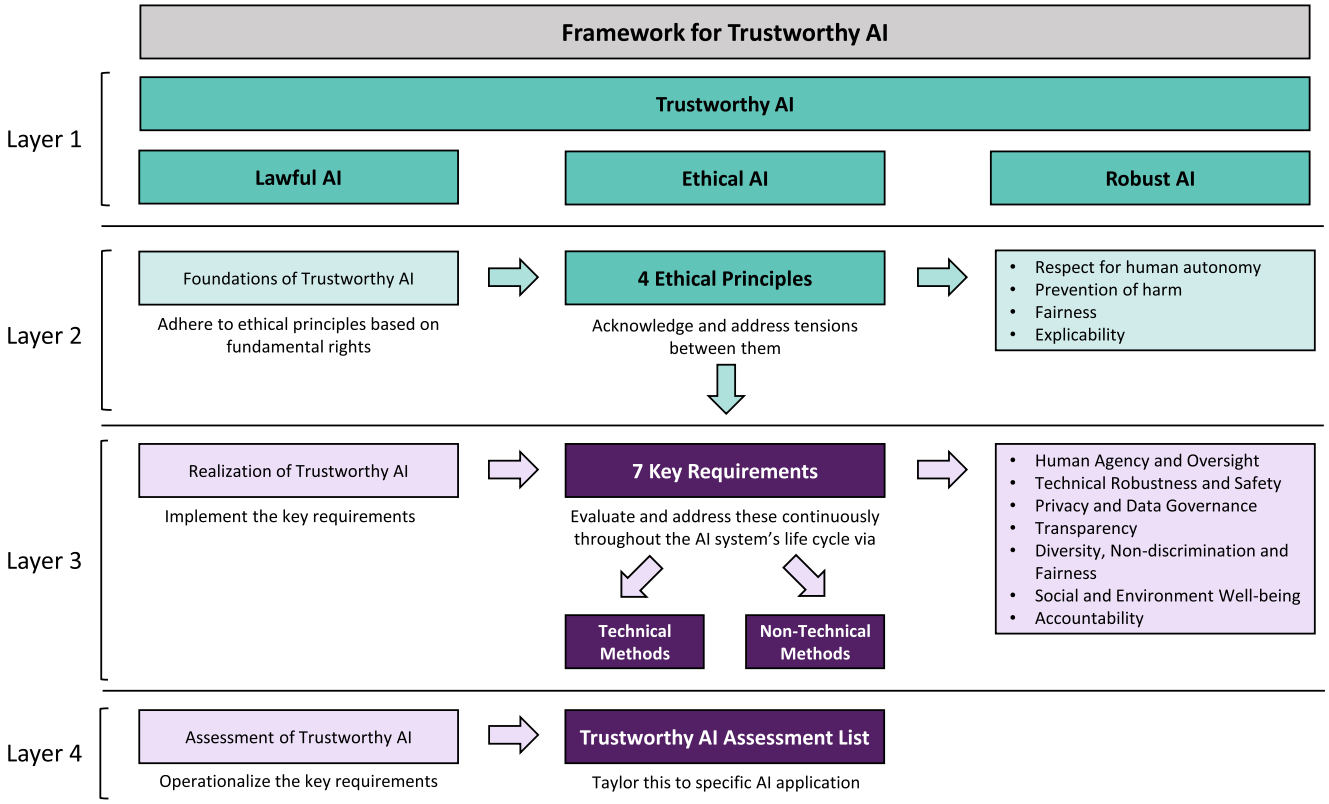

FIGURE 3. The guidelines as a framework for trustworthy AI.

is no fixed solution. There may also be situations where no ethically acceptable trade-offs can be identified.

2) REALIZING TRUSTWORTHY AI

The principles are translated into seven concrete requirements applicable to different stakeholders: developers, deployers and end-users, and the broader society:

1) **Human Agency and Oversight;**
2) **Technical Robustness and Safety;**
3) **Privacy and Data Governance;**
4) **Transparency;**
5) **Diversity, Non-discrimination and Fairness;**
6) **Societal and Environmental Well-being;**
7) **Accountability.**

All requirements are of equal importance, see Figure 4. Context and potential tensions between them need to be considered [26]. The implementation of these requirements should take place throughout the life cycle of an AI system [27]. Both technical and non-technical methods can be used to implement the above requirements [28]. These cover all stages of the life cycle of an AI system (continuous process) [29].

3) ASSESSING TRUSTWORTHY AI

Based on the seven key requirements, a non-exhaustive Trustworthy AI Assessment List [30] is provided to

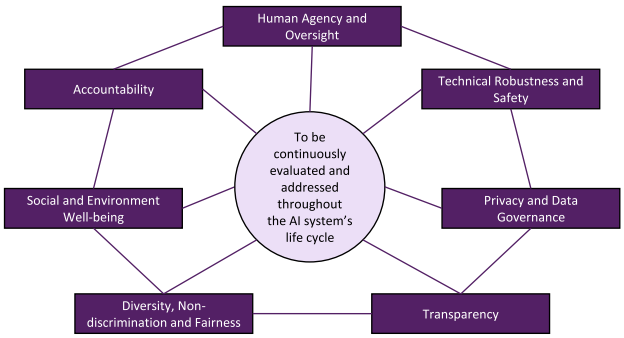

FIGURE 4. Requirements of trustworthy AI: Interrelationship of the seven requirements - all are of equal importance, support each other, and should be implemented and evaluated throughout the AI system's lifecycle.

operationalize Trustworthy AI, see Figure 5. This assessment list helps to evaluate whether the AIS that is being developed, deployed, procured or use the seven requirements of Trustworthy Artificial Intelligence (AI) as specified in the Ethics Guidelines for trustworthy AI. This Assessment List for Trustworthy AI (ALTAI), as detailed by [20], is crafted for self-assessment purposes, offering an initial framework for evaluating trustworthy AI. Designed for flexibility, organizations can adapt elements from ALTAI

that are pertinent to their specific AI system or incorporate additional components based on their sector of operation. ALTAI aids in comprehending the essence of trustworthy AI, delineating potential risks associated with an AI system, and providing guidance on minimizing these risks while maximizing AI benefits. The primary purpose of ALTAI is to assist organizations in understanding the potential risks posed by proposed AI systems and to determine proactive measures required to mitigate those risks. It serves as a comprehensive tool to identify how AI systems may introduce risks and guides organizations in implementing measures to prevent and minimize these risks, thereby ensuring a balanced optimization of AI benefits.

For comparison, the scientific AI community often uses very similar principles to describe trustworthy AIS [31], [32]:

- Ensure safety - establish accountability,
- Ensure fairness - uphold human rights and values,
- Respect privacy - reflect diversity/inclusion,
- Promote collaboration - avoid concentration of power,
- Provide transparency - acknowledge legal/policy implications,
- Limit harmful uses of AI - contemplate implications for employment.

B. THE CONCEPT OF ODD AND ITS RELATED BEHAVIOR COMPETENCIES

1) OPERATIONAL DESIGN DOMAIN (ODD)

ODD is a concept for specifying the operating conditions of an automated system, often used in the field of Autonomous Vehicles (AV). These operating conditions include environmental, geographical and time of day constraints, traffic, and roadway characteristics [33]. At the highest level, the ODD is divided into the following attributes [33]:

- Scenery elements
- Environmental conditions
- Dynamic elements

The scenery elements attribute consists of the spatially fixed elements of the operational environment (e.g., roads, traffic lights, etc.) relative to the ego vehicle (in terms of the position of the elements). The environmental conditions attribute includes weather and atmospheric conditions (including connectivity). The dynamic elements attribute describes moving elements of the ODD, e.g., traffic, subject vehicle, etc. Figure 6 illustrates a top-level taxonomy of ODD attributes. All attributes are considered to be of equal importance [34].

The concept of ODD indicates that automated systems have limitations and should operate within predefined restrictions to ensure safety and performance [35]. The definition of an ODD is crucial for developers and regulators to establish clear expectations and communicate the intended operating conditions of automated driving systems (ADS). An ADS is defined as the hardware and software collectively capable of performing the full dynamic driving task (DDT) on a sustained basis, regardless of whether it is limited to an ODD. The DDT encompasses the necessary tactical and operational functions required to operate the ADS-equipped vehicle, including the interdependent categories of sensing and perception, planning and decision, and control [36]. In summary, an ODD defines the operating conditions under which an ADS is designed to operate safely. However, the Target Operational Domain (TOD) comprises the real-world conditions that an ADS may experience during its deployment. While the ODD consists of the operating conditions in which an ADS is designed to operate, the TOD is the area in which the ADS is deployed and may include conditions outside the ADS's ODD. Often, the Operational Domain (OD) will generally be a superset of the ODD properties, representing real-world conditions that an ADS may encounter. In real-world use of an ADS, the difference between an ODD and a TOD highlights the limitations of the ADS. In all practical cases, an ODD definition will not be exhaustive enough to cover all the attributes or occurrences in a TOD. Therefore, it is crucial to objectively define the boundary between ODD and TOD and to incorporate design mechanisms in the ADS to execute fallback maneuvers when an ODD exit is encountered, ensuring safe operation in a TOD. The current operational domain (COD) refers to the real-time operational domain, i.e., the real-time real-world conditions experienced by the ADS (Figure 7). Consequently, the COD can vary for every time instance. The relationship between the COD and the OD is described by the following equation:

$$OD = \max_i COD(t_i) \tag{1}$$

Therefore, the maximum COD across all time instances is equal to the OD.

As a result of the introduced ODD concept, the remaining residual risks, reflecting the limitations of the ADS, related to the deployment and the current operation can be quantified as the difference between the TOD or COD and the ODD for which the ADS was designed, see Figure 7.

2) BEHAVIOR COMPETENCIES (BC)

Originally, the term "behavior competency" refers to measuring and assessing human performance in specific applications, such as evaluating navigational competence [37] or understanding factors influencing high-performing system engineers [38]. The versatility of this concept in categorizing human performance makes it an intriguing candidate for describing the capabilities of AIS. In the ADS domain, a specific set of BCs enables the AV to operate safely within the TOD. Each TOD necessitates a mandatory set of BCs to comply with traffic rules and handle interfering ODD attributes. In essence, a behavior refers to goal-directed actions taken by an engaged ADS during the DDT or DDT fallback within the ODD (if applicable) at a variety of timescales [39]. Specifically, a BC is defined as the expected and measurable capability of an ADS function to operate a vehicle within its ODD. Here, competence encompasses the

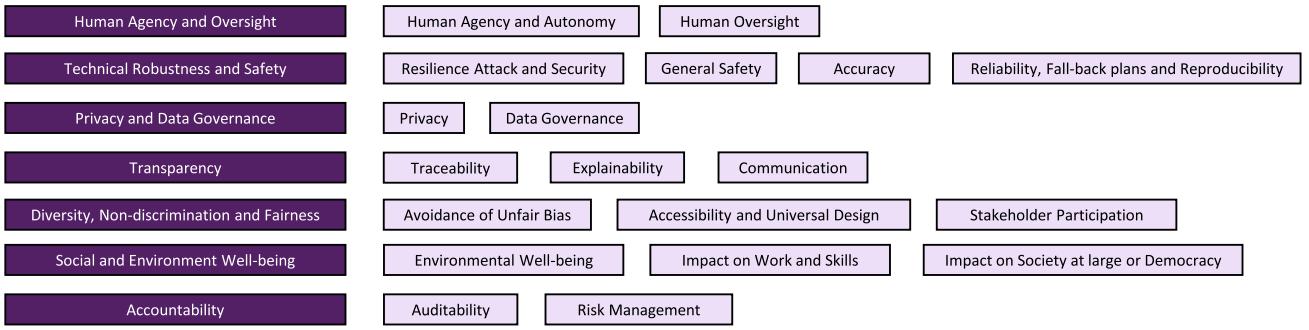

FIGURE 5. Seven key requirements and their related sub-requirements implemented in the final assessment list for trustworthy AI (ALTAI).

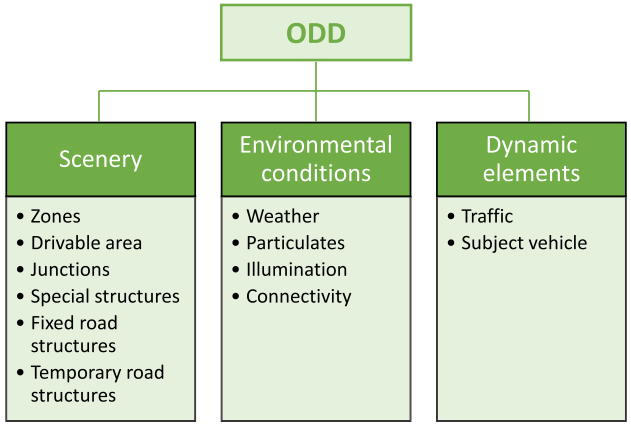

FIGURE 6. Top level taxonomy with ODD attributes.

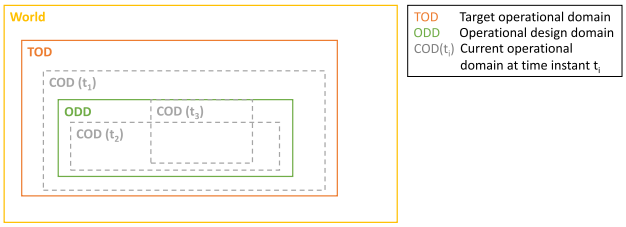

FIGURE 7. Relationship between ODD, TOD and COD.

term 'expected' in the definition. Leveraging skills, knowledge, and abilities, an ADS competently performs behaviors according to criteria established by the ADS developer [39]. The BC of an ADS can be identified using analytical frameworks suggested by [40] and [41]. These frameworks encompass an ODD and driving situation analysis, followed by an Object and Event Detection and Response (OEDR) analysis. The OEDR analysis involves identifying relevant objects and events necessary for detection, followed by determining the respective response, constituting elementary BCs.

3) AUTOMATED TRANSPORT SYSTEMS (ATS)

The same concept applies to ATS in general, encompassing land, air, and marine operations, such as autonomous trains, drones, ships, etc. [42], [43]. As with the ADS, the initial step towards assurance involves developing an accurate and clear understanding of the requirements. A crucial aspect of formulating requirements for an ATS is comprehending its operating conditions and behavioral capabilities. In the transport domain an ODD definition includes all static, dynamic and environmental attributes like weather and connectivity [42], [43]. Ensuring completeness of requirements from an ODD perspective necessitates capturing a wide variety of actors and their diverse characteristics in the ODD definition. This includes actor attributes such as type, skin color, disabilities, etc. It's important to note that an ODD definition doesn't specify the behavioral capabilities or the desired behavior of the system. The behavioral capabilities of automated systems refer to the skills and characteristics that an ATS should possess to navigate and interact with the environment effectively and safely. These capabilities are designed to ensure that ATS can maneuver, make decisions, and respond appropriately to various situations on land, in the air, and at sea. Consequently, the ODD and its associated behavioral capability definition collectively form the system concept for an ATS.

4) ROBOTS

In addition to the transport domain, the concept of ODD and related behavioral competencies is also widely used in the robotics domain, e.g. agricultural robots [44] and other robots in the field of production and automation.

5) ARTIFICIAL INTELLIGENCE-BASED SYSTEMS (AIS)

Having successfully applied the ODD concept to ADS-equipped vehicles, ATS and robotics to specify the operating conditions under which these systems operate safely, along with their limitations and corresponding constraints, the next logical step is to extend this concept to AIS. This intension represents the progression from its origin in ADS through ATS to the most general domain of AIS, as illustrated in Figure 8. Undoubtedly, this extension is challenging, as it introduces many new application domain-specific aspects into the description of operating conditions. However, it is also promising, given that ADS and ATS already incorporate

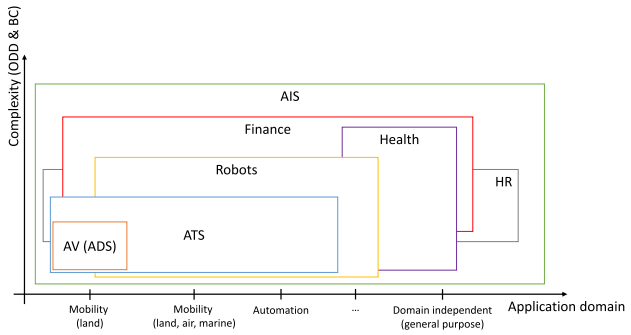

FIGURE 8. Increasing complexity of ODD and its related BC along the application domain.

numerous AI elements. It is noteworthy that ADS and ATS are inherently AI-driven, as many core functionalities in the well-known sense-plan-act cycle that these systems operate on are based on AI approaches and algorithms. As the ODD concept expands into the AIS domain, it becomes a comprehensive framework for characterizing the operational landscape of diverse intelligent systems, incorporating the evolving complexities of AI-driven applications.

6) CONCLUSION

ODD and its related behavior competencies play a crucial role in establishing system boundaries for high-risk systems, including ADS, ATS, and robots. This framework enables the precise definition of operational limits, providing a foundation for assessing and managing risk. Currently, one significant source of uncertainty in the output of AIS, such as ADS, ATS, and robots, is scope compliance. Scope compliance involves evaluating whether models are used within the intended scope, based on training and test data. To address this, a target application scope (TAS) for the respective AIS can be defined [45]. This TAS can be viewed as analogous to the ODD in the context of AIS. However, the TAS concept has limitations, particularly in conducting a comprehensive scope compliance assessment, as it does not encompass behavior competencies. Therefore, it is reasonable to apply the ODD and behavior competencies concept to AIS, as this approach provides a more comprehensive framework for assessing and ensuring the appropriate scope and behavior of intelligent systems.

C. RISK ASSESSMENT AS A CENTRAL ELEMENT

Assessing the risk of an AIS is a complex task due to the vast array of subtasks involved, such as data collection and verification, feature extraction, and monitoring [46]. Currently, various risk assessment frameworks are emerging [47], [48]. These frameworks either focus on qualitative aspects, such as question banks [49], or quantitative assessments [50]. In the quantitative approach, the focus is on the desired properties of quantitative assessment metrics, along with specific aspects like uncertainty estimation, error propagation, and out-of-distribution detection [51]. Consequently, the entire AI process revolves around risk, spanning from classification to assessment and culminating in appropriate management, facilitating the deployment of trusted AIS. The concept of ODD and its related BC can act as a central approach for efficient and effective risk analysis. Similar approaches have shown great promise in the field of ADS, where identifying residual risks is a key deployment measure [52]. Safeguarding AIS, which ensures safe operation, can be formulated as an overarching goal of achieving the absence of unreasonable/unacceptable risk (AUR) [53], [54], [55], [56]. In other words, the risk needs to be reduced to an acceptable level. Establishing a consensus within society is crucial for determining the acceptance threshold. Arguing for the absence of AUR underscores the necessity of risk quantification. The risk assessment of high-risk classified AIS, quantifying acceptable residual risks for deployment, represents a central element in realizing human-centric and trustworthy AIS' in the EU market.

1) DEFINITION OF RISKS ALONG THE AI LIFECYCLE

Risk and its assessment have been fundamental in the automotive functional safety domain for an extended period, with well-established and approved practices in various safety-critical automotive use cases. ISO 26262, for instance, defines risk as a combination of the probability of harm occurrence and the severity of that harm [57]. Similar definitions are present in ISO/IEC 25010 [58] and UL 4600 [59]. The term "residual risk" is also well-established in different norms, particularly within the automotive domain. ISO 26262 defines residual risks as the risks that persist after the implementation of safety measures. This term is further utilized in the standard ISO/PAS 21448 [60], which focuses on the safety of the intended functionality (SOTIF) in road vehicles. SOTIF is defined as *"the absence of unreasonable risk due to hazards resulting from functional insufficiencies of the intended functionality or by reasonably foreseeable misuse by persons."* Adopting a risk-based approach, limitations of the AIS result in residual risks during deployment and operation within a specific ODD. Residual risks can be categorized across the AI life cycle based on when they originate and become relevant—spanning from the design and development phases to deployment, marking the pre-deployment phase (refer to Figure 9). After successful AIS deployment, the post-deployment phase commences, involving operational use, monitoring, and concluding with the evaluation and analysis phase. In this context, all phases along the AI life cycle are intricately connected to risks. The circular nature of the AI life cycle underscores the importance of reevaluating and assessing all pertinent risks.

2) RISK TYPES AND THEIR ORIGINS

Concrete risks for AIS emerge from malicious use, competitive pressures, and organizational factors [61]. By defining AIS, one can categorize risks into those associated with the

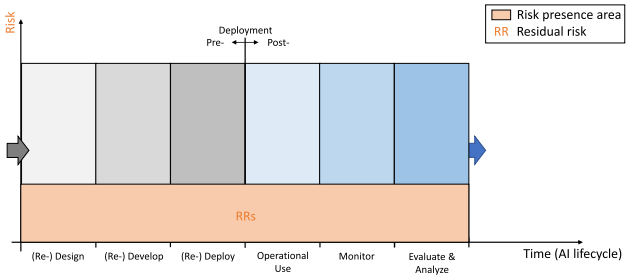

FIGURE 9. Residual risks along the AI lifecycle.

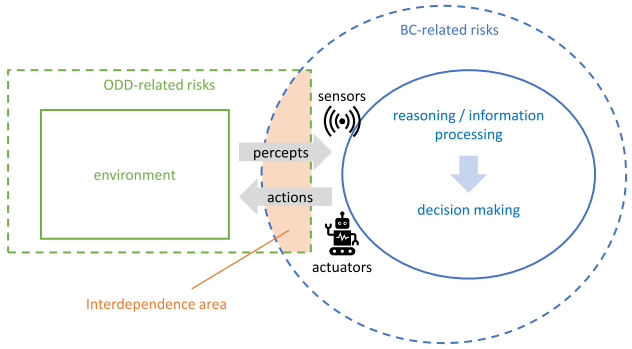

FIGURE 10. ODD and BC related risk origin and their interdependence.

environment in which the AIS operates and those inherent to the AIS itself, encompassing all AI elements such as hardware and software. These elements are essential for perceiving the current environment, making decisions, and implementing actions through actuators. In this context, risks related to the environment are quantified via the ODD concept, while risks tied to the AIS itself are described via the related BC concept. Consider the following example from the ADS domain: the elementary behavioral competency of "lane change" is part of the DDT for an ODD covering both motorways and urban roads. However, the risk associated with executing a lane change may significantly differ within this ODD. It is evident that ODD and BC are inherently intertwined, as they continuously influence each other on a structural basis. This interdependence is illustrated in Figure 10.

Following that rationale, the following risk types linked to the ODD and BC concept can be classified along the AI lifecycle, including the pre- and post-deployment phase:

- **Boundary residual risks (BRR)** related to ODD and BC boundaries originating in the design phase:

$$BRR = BRR_{ODD} + BRR_{BC} \tag{2}$$

- **Deployment residual risks (DRR)** related to the ODD and BC quantifiable in the deployment phase:

$$DRR = DRR_{ODD} + DRR_{BC} \tag{3}$$

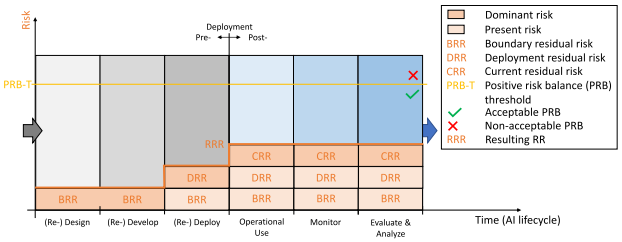

FIGURE 11. Dominant and present residual risks along the AI life cycle resulting in an acceptable/non-acceptable PRB evaluation.

- **Current residual risks (CRR)** popping up via the start of the operational use:

$$CRR = CRR_{ODD} + CRR_{BC} \tag{4}$$

- **Risk acceptance criteria (RAC)** being present during the entire AI lifecycle using the concept of positive risk balance (PRB) to be updated in every evaluation and analyzation phase.

Figure 11 provides a detailed depiction of classified risks along the AI lifecycle. It is crucial to emphasize that none of the emerging risks will disappear throughout the AI lifecycle; rather, their dominance, importance, or role will undergo changes, as indicated by the comparison between dominant and current risks. Consequently, the resulting RR risk can be quantified as the sum of individual risks. The primary objective throughout the AI lifecycle is to maintain all residual risks below a predetermined threshold (indicated by the area with a green tick), representing an acceptable Probability of Residual Risk (PRB) evaluation result, denoted as PRB-T. Any breach of this threshold (marked by the area with a red cross) results in an unacceptable residual risk, prompting immediate implementation of appropriate countermeasures.

3) RESIDUAL RISKS RELATED TO THE ODD CONCEPT

4) ODD BOUNDARY RESIDUAL RISK (ODD-BRR)

The ODD-BRR quantifies the residual risk associated with the boundary of the ODD to be certified/homologated, see Figure 12. The boundary can be determined by two different approaches: First, from the inside, by drawing the boundary based on the covered ODD segments, resulting in a coherent ODD with a defined boundary. Second, an outward approach, by clustering dysfunctional cases to create a boundary named as safety boundary SB_{OD} based on where the AIS doesn't behave as expected. In this approach, the boundary is determined by applying a safety margin to the dysfunctional area characterized by SB_{OD}, resulting in the ODD boundary itself. Both approaches carry the risk of setting a boundary that may be either overly conservative or excessively ambitious. The accuracy of the inside approach relies on the underlying assumptions used to classify an ODD segment as covered. For the outside approach, accuracy is contingent on the applied safety margin to the

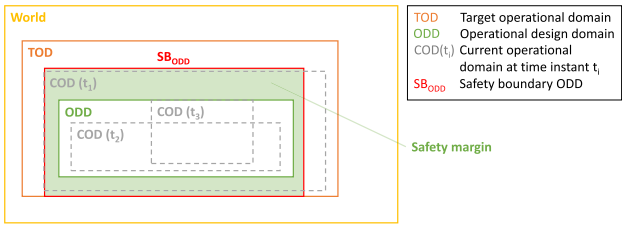

FIGURE 12. The concept of ODD enables to define residual risks related to the boundary, the deployment and the current situation.

dysfunctional area.

$$BRR_{ODD} = \frac{(TOD - ODD)}{\text{Safety Margin}} \tag{5}$$

Depending on the used apporach to determine the boundary (inside or outside) the safety margin reads as:

$$\text{Safety Margin} = \begin{cases} SB_{ODD} - ODD, & \text{outside approach} \\ 1, & \text{inside approach} \end{cases} \tag{6}$$

5) ODD DEPLOYMENT RESIDUAL RISK (ODD-DRR)

The ODD-DRR describes the residual risk at the stage of the deployment of the AIS. It is quantified as the difference between the TOD and the covered ODD after the AIS V&V phase, see Figure 12. In that sense, the uncovered parts of the ODD represent the residual risk present at the deployment time.

$$DRR_{ODD} = TOD - ODD \tag{7}$$

6) ODD CURRENT RESIDUAL RISK (ODD-CRR)

The ODD-CRR defines the current residual risk during the operation of an AIS. It is quantified as the difference between the COD at time instance t_i and the covered and therefore certified/homologated ODD after the deployment phase, see Figure 12. In that sense, the CRR represents an important post-deployment risk measure. During operation, the COD can be (partly) outside the certified/homologated ODD. After operationalization the COD can be inside the certified/homologated ODD as well as outside. In both cases, the reaction of the AIS considering the CRR is essential for enabling trustworthy AIS.

$$CRR_{ODD} = COD(t_i) - ODD \tag{8}$$

7) RESIDUAL RISK RELATED TO THE BC CONCEPT

In addition, residual risks can be linked to the BC of the targeted ODD. In that sense, limitations of the BC related to the ODD of the AIS again result in residual risks for the deployment and operation. Again, four types of residual risks can be distinguished.

8) BC BOUNDARY RESIDUAL RISK (BC-BRR)

The BC-BRR quantifies the residual risk associated with the boundary of the BC that requires certification or homologation, as depicted in Figure 13. Similar to the ODD concept, two distinct approaches can be employed to determine the boundary:

- Inside Approach: This method involves drawing the boundary from the inside, based on the approved BC features, resulting in a cohesive BC with a well-defined boundary.
- Outward Approach: In this approach, dysfunctional cases are clustered to form a boundary referred to as the safety boundary SB_{BC}. This boundary is established where the AIS deviates from expected behavior. The BC boundary is then derived by applying a safety margin to the dysfunctional area characterized by SB_{BC}.

Just like in the case of the ODD concept, both approaches carry the risk of setting a boundary that may be either overly conservative or excessively ambitious. Using only the approved BC features may result in a boundary that is too conservative or too ambitious, depending on the underlying assumptions used to classify a BC feature as covered. On the other hand, applying too large a safety margin to the dysfunctional area will result in a boundary that is too conservative, while applying too small a safety margin will result in a boundary that is too ambitious.

$$BRR_{BC} = \frac{(TBC - BC)}{\text{Safety Margin}} \tag{9}$$

Depending on the used approach to determine the boundary (inside or outside) the safety margin reads as:

$$\text{Safety Margin} = \begin{cases} SB_{BC} - BC, & \text{outside approach} \\ 1, & \text{inside approach} \end{cases} \tag{10}$$

9) BC DEPLOYMENT RESIDUAL RISK (BC-DRR)

The BC-DRR characterizes the residual risk at the stage of deploying the AIS. It is quantified as the TOD normalized by the intersection of the TOD and the associated BC, see Figure 13. In this context, the smaller the mandatory overlap between the TOD and the associated BC, the higher the residual risk will be. A reduced overlap signifies a limited portion of BC that is available to safely operate within the TOD. Put differently, a smaller portfolio of BC increases the likelihood of encountering critical situations during operation, leading to higher residual risks.

$$DRR_{BC} = \frac{TOD}{\text{Intersection}(TOD, BC)} \tag{11}$$

10) BC CURRENT RESIDUAL RISK (BC-CRR)

The BC-CRR defines the current residual risk associated with the BC during the operation of an AIS. It is quantified via the COD at time instant t_i, normalized by the intersection of $COD(t_i)$ and the associated BC, as depicted in Figure 13. Once again, the BC-CRR stands as a crucial post-deployment risk measure. During operation, the COD may extend (partly) beyond the certified/homologated ODD, significantly impacting the necessary BC to operate within the COD. In both cases, the reaction of the AIS, considering

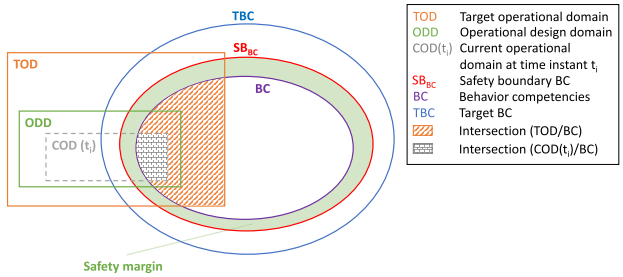

FIGURE 13. Residual risks related to the BC boundary, the deployment and the current situation.

the BC-CRR, is essential for enabling a trustworthy AIS.

$$CRR_{BC} = \frac{COD(t_i)}{\text{Intersection}(COD(t_i), BC)} \tag{12}$$

11) RISK ACCEPTANCE CRITERIA (RAC)

Introducing AUR as an overall safety goal prompts the question of when this goal is sufficiently achieved. Various risk acceptance criteria can provide guidance in this regard. Examples include ALARP (As Low As Reasonably Practicable), GAMAB (Globalement Au Moins Aussi Bon - a fixed target stipulating that new technical systems should be at least as safe as comparable ones), and PRB (Positive Risk Balance). PRB comes from the ADS domain and refers to the "Benefit of sufficiently mitigating residual risk of traffic participation due to automated vehicles" [53], [62]. When employing the PRB approach, a key question arises: how can the threshold be determined in a structured, explainable, and accepted manner? Estimating AUR (and thus PRB) is primarily a pre-deployment task, making the determination of a meaningful threshold challenging. One potential approach involves using exposure to risk as a guiding factor. In the post-deployment phases, the absence of AUR (and thus PRB) can be measured, particularly in the ADS domain through in-service monitoring. This enables a more feasible setting of a threshold for risk acceptance criteria.

12) RISK MITIGATION

Mitigating the associated risk of an AIS is imperative throughout the entire AI lifecycle. While the current emphasis is on mitigation actions during the development phase to reduce the risk at deployment, concrete strategies involve formal design and analysis of systems incorporating AI components [63]. Probabilistic programming languages are also emerging for such tasks [64]. In the pursuit of robust interpretability, [65] employs worst-case interpretation and introduces a probabilistic notion of robustness. Furthermore, [66] addresses risk mitigation through formal certification but focuses exclusively on one aspect of the AI lifecycle. However, post-deployment phases introduce several degrading factors, including failures, faults, defects, misbehaviors, maintenance needs, unknown dysfunctional cases, new restrictions, etc., which may elevate residual risk beyond acceptable limits. In such cases, reducing the TOD and/or BC can help to bring the residual risk back below acceptable

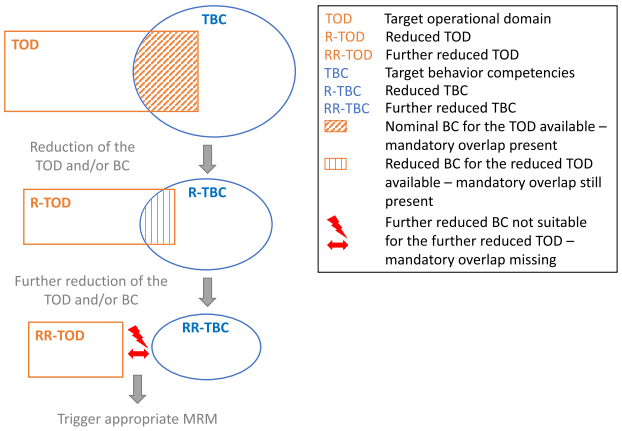

FIGURE 14. Residual risk reduction limitations. Mandatory overlap of TOD and TBC decides if the operation of the AIS can be continued, in case not, a safe termination of the operating AIS has to be initiated, e.g. via an appropriate MRM.

limitations, as illustrated in Figure 14. This approach is viable as long as the mandatory intersection between TOD and TBC ensures safe and reliable AIS operation within its ODD, leveraging the embedded TBC. This intersection underscores the inherent interdependence between the ODD and its associated BC, forming the foundation for the trustworthy operation of any AIS. Should the mandatory intersection be absent, the operation of the AIS must be halted in a safe manner. For example, in ADS', this could involve triggering a corresponding Minimum Risk Maneuver (MRM) in such situations.

III. METHODOLOGY

The overarching goal of our approach is to ensure the trustworthy behavior of AIS throughout the entire AI lifecycle. The proposed methodology hinges on a systematic transfer of the ODD concept and its associated BC to AIS, facilitating a structured development and pre-development phase of AIS. Specifically, this strategy incorporates a Trustworthiness Assurance Process that addresses the central elements of the Trustworthiness Assurance Framework. It enables Trustworthiness Argumentation that considers the overall objective while addressing all relevant sub-objectives related to trustworthy AIS. Importantly, this process seamlessly integrates with the entire AI lifecycle, meeting a crucial prerequisite. The process views ODD and BC elements as the primary sources of requirements to be justified throughout the AI lifecycle, characterizing trustworthy AIS. This intended process sets the stage for a traceable certification of AIS based on ODD/BC elements for specific target application areas, accounting for use restrictions. Moreover, it establishes a structured path for re-certification purposes based on this approach.

A. DESCRIPTION OF THE TARGET APPLICATION AREA VIA ODD AND BC ELEMENTS

The concept of ODD and its associated BC serves as a promising tool for efficiently and effectively describing the

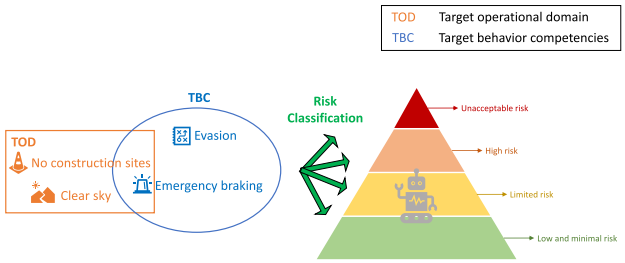

FIGURE 15. AIS risk classification based on TOD and its related TBC.

intended target application area, incorporating all relevant restrictions of use. This directly addresses a prevalent human concern about AI—namely, the fear of losing control over AI solutions and being unable to dictate their actions, especially with the increasing focus on general-purpose AI that extends to military purposes [67]. From a technical standpoint, the ODD and BC elements meticulously outline the current limitations of AIS through their ODD and BC boundaries (refer to section II-B). Established and widely accepted standards play a critical role in describing the target application area, preventing misuse and misunderstanding across the entire AIS value chain and beyond. This allows restrictions of use to be communicated in a traceable and understandable manner, aligning with the human-centered principles of the AI Act. Equally significant is the fact that violations of the target application areas can be easily measured and monitored throughout the lifecycle, given that standardized ODD and BC elements provide a reference basis. Additional details on standards are covered in section V. In alignment with the central element of the AI Act, the risk-based approach, the pyramid of risks is employed to classify AIS into four concrete levels of risk. The ODD and BC elements support a traceable and consistent classification of the AIS system into unacceptable risk, high risk, limited risk, and low or minimal risk, significantly reducing interpretation uncertainty. This qualitative representation is illustrated in Figure 15, addressing another major concern for companies unsure about how their AIS will be classified. In essence, incorporating the ODD and its associated BC elements throughout the AI lifecycle ensures the consistency, traceability, and acceptability of the entire operation of the AIS.

B. TRUSTWORTHINESS ASSURANCE FRAMEWORK AND RELATED PROCESS

Current assurance activities regarding AIS predominantly focus on machine learning (ML), considering ML systems as a subset of AIS with reduced ODD and BC, simplifying the task of trustworthiness assurance. However, compared to assurance processes for conventional systems, ML systems often grapple with lackluster requirements leading to unclear specifications [68]. In that context, the concept of responsible research and innovation (RRI) has a long history and evolved into a framework for the entire research, development and innovation lifecycle [69]. In constrast, the emerging field of responsible AI (RAI) has more research output, but the uptake of RI by RAI is slow, showcasing an independent development of those two concepts [70]. With current EU initiatives towards RRI and the fact that ML is lagging behind in assurance topics compared to e.g., the ADS domain, approaches that use existing concepts and frameworks and apply it towards AIS are urgently needed.

Advancements have been made in safety assurance cases for machine learning components, taking requirements, data management, and model learning into account. Yet, these approaches are currently limited to specific machine learning implementations tailored for particular use cases [71]. Robustness, monitoring, steering of ML systems, and hazard reduction for deployment are identified as primary challenges for ML systems in terms of safety [72], [73]. Activities around trustworthiness assurance and validation, especially for ML systems, are currently more prevalent but are often performed at a small scale and lack comprehensive coverage across the complete lifecycle [74]. However, emerging frameworks for AI validation, such as those in medicine [75], [76], suggest a growing recognition of the need for comprehensive validation practices. For AIS beyond ML systems, addressing safety challenges by design becomes increasingly crucial due to higher complexity and opacity, potentially leading to elevated risks [77]. Continuous safety assurance processes for ML systems are actively under investigation, exploring the development of assurance cases in alignment with specific safety standards determined by safety integrity levels (SILs) [78], [79], [80].

The trustworthiness assurance framework represents a systematic and structured approach that serves as a foundation for designing and developing trustworthy AIS. Frameworks, in general, are designed to streamline and simplify development by offering a predefined structure and reusable components, aiming to save time, reduce complexity, and promote consistency. Figure 16 illustrates the further enhanced trustworthiness assurance framework, building upon the initial version shown in Figure 3. The primary objective of the framework is to provide a structured pathway to robust trustworthy reasoning that aligns with all seven requirements (depicted in purple) and their associated sub-requirements (depicted in light purple), as specified in the AI Act. It's crucial to note that the method employed (e.g., ALTAI, FMEA, etc.) and the evidence presented will vary based on the current target application area. The adaptability of the framework allows it to cater to diverse contexts, ensuring a tailored and effective approach to trustworthiness assurance.

Based on the outlined trustworthiness assurance framework, a corresponding process, termed the trustworthiness assurance process, is designed to furnish robust and scalable evidence for the overarching goal of trustworthy argumentation. This goal encompasses all relevant sub-goals related to ensuring the trustworthiness of AIS. Illustrated in Figure 17, the high-level trustworthiness assurance process

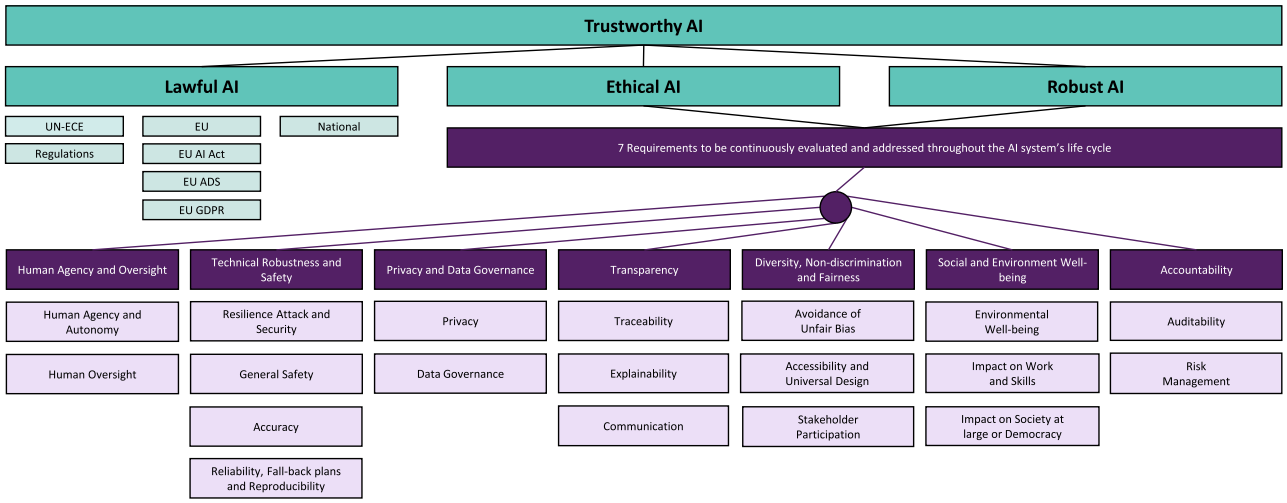

FIGURE 16. Trustworthiness assurance framework for AIS.

comprises six major steps, strategically timed to align with the development phases of AIS. The process is designed to be executed iteratively, allowing for updates in terms of ODD/BC extensions and/or reductions. A similar argumentation-generating process for safety assurance is actively explored across multiple domains, such as the ADS domain [81] and the domain of unmanned aerial vehicles [43].

Bringing in concepts and frameworks that are established in the ADS domain (standardized, specified, …) and apply it towards AIS assurance are urgently required. An enhancement towards trustworthiness by considering a broader scope of aspects, depending on the application domain, can thus be achieved. Furthermore, the practical significance of the suggested trustworthiness assurance assessment is the binary decision for or against the assurance.

Specifically, the six key steps within the Trustworthy Assurance Process relate to

1) Specification of the target application area using ODD and BC elements. This specification can be used to determine restrictions of use, which represent sub-goal 1. In addition, the defined ODD and BC elements form the basis for the AIS related risk classification and act as a major source of requirements for the AIS.
2) Specification of disturbance and/or fault injection campaigns to prove the robustness properties of the AIS against specific disturbances and/or faults. This step is directly linked to the trustworthiness argumentation formulated in sub-goal 2.
3) Data set and scenario preparation includes the specification of the benchmark scenarios and data set elements to be used for the evaluation of the AIS. This element includes the interface to the scenario databases, including scenarios related to the intended deployment area and specific parts of the datasets to be used for verification and validation purposes, e.g., k-fold approach. In addition, a mixture of synthetic and real data may be used during the verification and validation activities. Step 3 completes all the preparations needed to run/execute the AIS in Step 4. In this respect, step 3 relates directly to sub-target 4.
4) Running the AIS under test includes the complete evaluation of the AIS based on the specified Steps 1-3 together with all logging activities required to feed the relevant data into Step 4 to evaluate the corresponding KPIs and metrics.
5) The KPIs and Metrics process step focuses purely on the evaluation of the given KPIs and metrics needed to prepare the subsequent trustworthy argumentation. The KPIs and metrics are defined in such a way that

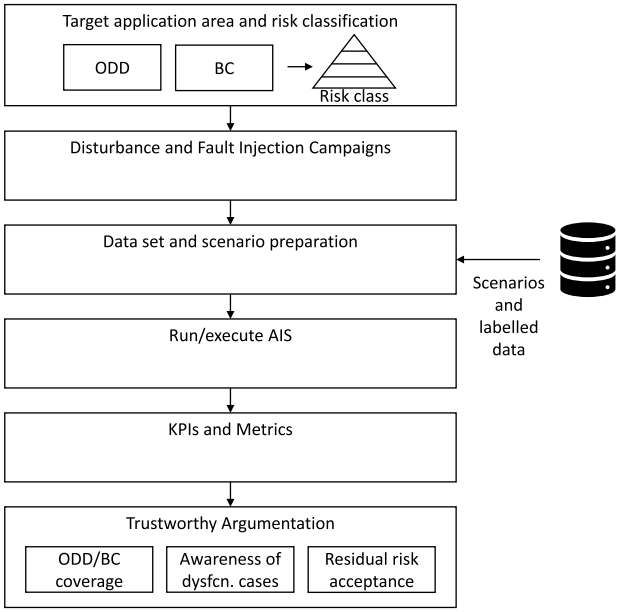

FIGURE 17. High-level trustworthy assurance process for AIS.

they measure the core characteristics associated with trustworthy AIS. Specifically, all seven requirements and related sub-requirements are translated into specific KPIs and metrics to be evaluated based on the constraints set in steps 1-3. Step 5 is directly linked to sub-goal 5.

6) Step 6 forms the trustworthy argument itself, which consists of three disciplines: first, the covered ODD/BC elements are evaluated with respect to the targeted application area; second, the awareness of so-called dysfunctional cases is evaluated; and finally, the residual risk is evaluated using appropriate risk acceptance criteria. A combination of all three disciplines establishes the basis for claiming a trustworthy AIS. Step 6 is the core element of sub-goals 2 and 3.

It is important to note that the entire process (including all six steps) relates directly to sub-goal 6, as all steps can be linked to the entire AI lifecycle, see Figure 18.

Figure 18 illustrates that all process steps are intricately linked to at least one specific phase within the AI lifecycle. Notably, several process steps, such as defining the target application area, establishing KPIs and metrics, etc., are linked to more than one AI lifecycle phase, emphasizing the significance of these process elements. Moreover, all process steps are present within both the pre- and post-deployment phases, underscoring the broad applicability of the process. This inclusivity allows for seamless updates in terms of ODD/BC extensions and/or reductions, reinforcing the adaptability and relevance of the process across the entire AI lifecycle. The assurance of any AIS can only be done if a certain level of trust is established. The required level of trust needs to be determined through this dedicated process which provides a binary assurance decision.

C. CERTIFICATION OF TRUSTWORTHY AIS

In general, certification refers to a formal process by which an AIS is assessed and verified to meet specific requirements outlined in the AI Act. It serves as a means of ensuring that a certain level of trustworthiness is achieved and maintained [82], [83]. In our case, the targeted Trustworthiness Assurance Process has the potential to play a significant role within the overall certification process. Certification of AIS serves several purposes, including

- Quality assurance: Ensuring that AIS meet pre-defined quality or performance standards.
- Compliance: Confirming that an organization adheres to specific regulations, standards, or industry best practice.
- Safety: Verifying that AIS meet safety requirements to protect consumers or the environment.
- Marketability: Enhancing the credibility and marketability of AIS by demonstrating compliance with recognized standards.
- Scalability: The formal process leading to a certified AIS has to be scalable to all types of AIS.

Certification is not a one-off process; it usually requires ongoing maintenance to ensure continued compliance [84]. Organizations may need to undergo periodic audits, retests or other assessments to renew their certification. In addition to ongoing maintenance, ODD/BC labels provide a promising and consistent solution to indicate for which ODD/BC attributes the AIS is certified. Renewal is always required whenever there is an ODD and/or BC extension and/or reduction from the last certification. In conclusion, certification offers several benefits, including

- Credibility: Certification provides evidence of compliance, which increases the confidence of consumers, customers, or partners [85].
- Competitive advantage: Certified organizations can gain a competitive advantage in the marketplace as certification can be a differentiating factor.
- Risk mitigation: Certification helps to reduce risks associated with quality, safety or regulatory non-compliance.

Certification is expected to play a crucial role in various AI-driven industries and sectors [86] by ensuring that products and services meet established standards, contributing to safety, quality, and reliability in a wide range of areas [87], [88], [89], [90].

IV. OVERALL GOAL: TRUSTWORTHY AND HUMAN-CENTRIC AIS

In this section, the overall goal of deploying trustworthy and human-centric AIS is broken down into seven related sub-goals (see Figure 19) which are discussed in detail in the following sections. The specified sub-goals are

1) Restrictions of Use
2) Trustworthy Assurance/Argumentation
3) Awareness of Dysfunctional Cases
4) Scenario Data Bases and Data Sets
5) Metrics and KPIs
6) AI Product Life Cycle
7) Human Factors

and address the main identified burning issues of the developed approach to assurance of trustworthiness.

A. SUB-GOAL 1: RESTRICTIONS OF USE

1) MOTIVATION

As discussed in section I, many AIS' are anticipated to be deployed in open context environments characterized by complexity and unpredictability. These characteristics raise concerns about the potential loss of human control over AIS actions and learning once deployed in such environments. Similar concerns and calls for regulation have been identified in the context of general-purpose AI models. Key characteristics of these models, such as their large size, opacity, and the potential to develop unexpected capabilities beyond their creators' intentions, have led to questions and ethical considerations. Studies on large language models (LLMs) like ChatGPT [91], [92], [93], [94] highlight ethical and social risks [95]. Despite efforts to mitigate these risks, LLMs, including GPT-4, still present challenges related to user safety, fundamental rights, and the generation of harmful and criminal content. Privacy concerns also arise

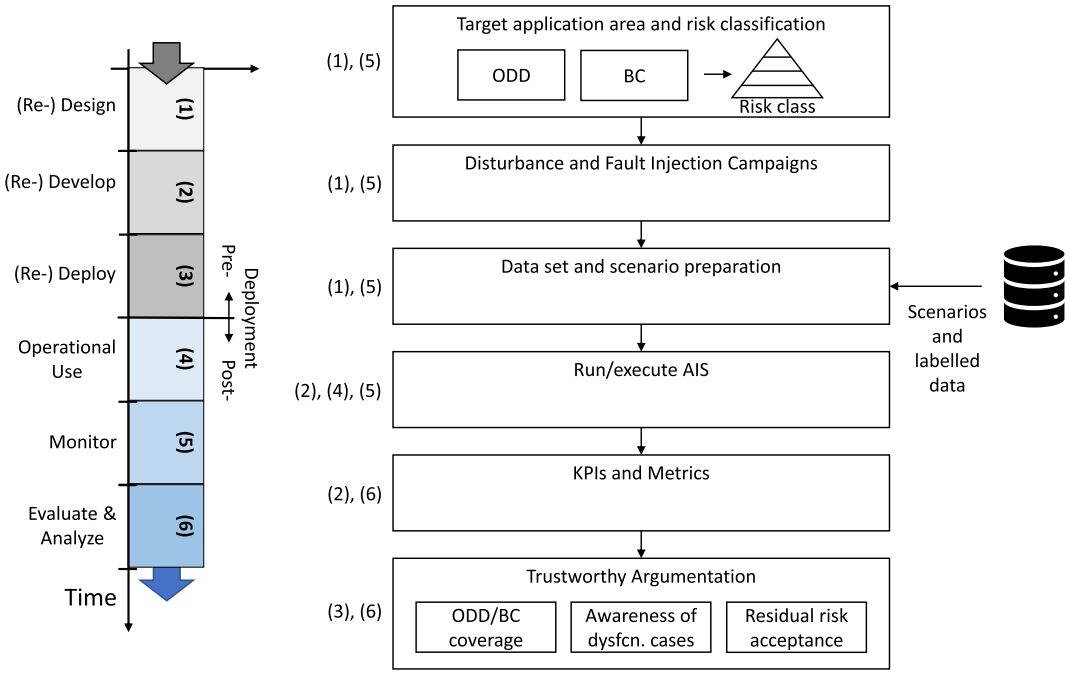

FIGURE 18. Trustworthiness assurance process linked to the AI life cycle.

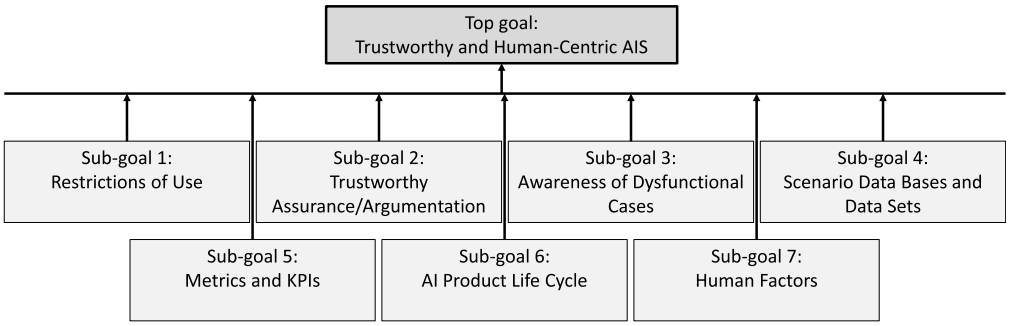

FIGURE 19. Top-goal breakdown into seven related sub-goals targeting trustworthy and human-centric AIS.

as general-purpose AI models are trained on publicly available data from the internet, raising issues related to plagiarism, transparency, consent, and lawful grounds for data processing. The question of liability for harm caused by general-purpose AI systems has been a topic of discussion. Calls for oversight and monitoring of AI through evaluation and testing mechanisms, as outlined in the AI Act [1], further emphasize the need for robust governance. In light of these challenges, innovative approaches are essential to describe the open context and/or general purpose of AI, including defining their boundaries and capabilities within the deployment environment. These approaches must address ethical, legal, and societal considerations to ensure responsible and accountable deployment of AI systems.

2) METHOD

The application of the ODD concept and its associated BC is pivotal in enabling a standardized, explainable, traceable, effective, and efficient risk classification of AIS' in accordance with the specified risk categories outlined in the AI Act. It is anticipated that ODD and BC elements will significantly enhance the risk classification process, surpassing the current focus on high-risk applications. The information content embedded in ODD and BC elements allows for the differentiation of various use cases within an application, leading to distinct risk classes. To address concerns regarding the fear that AI may learn tasks beyond its original design, the formulation of usage restrictions becomes essential. These restrictions are crucial for rebuilding human trust in AI and providing clear guidelines post-deployment regarding which ODD and BC elements the AIS can and should handle, with well-defined boundaries for non-legal and non-intended applications. Figure 20 illustrates a qualitative process for determining restrictions on the use of high-risk AIS. This process outlines the steps involved in establishing clear boundaries to mitigate risks and ensure responsible and accountable use of AI in high-risk scenarios.

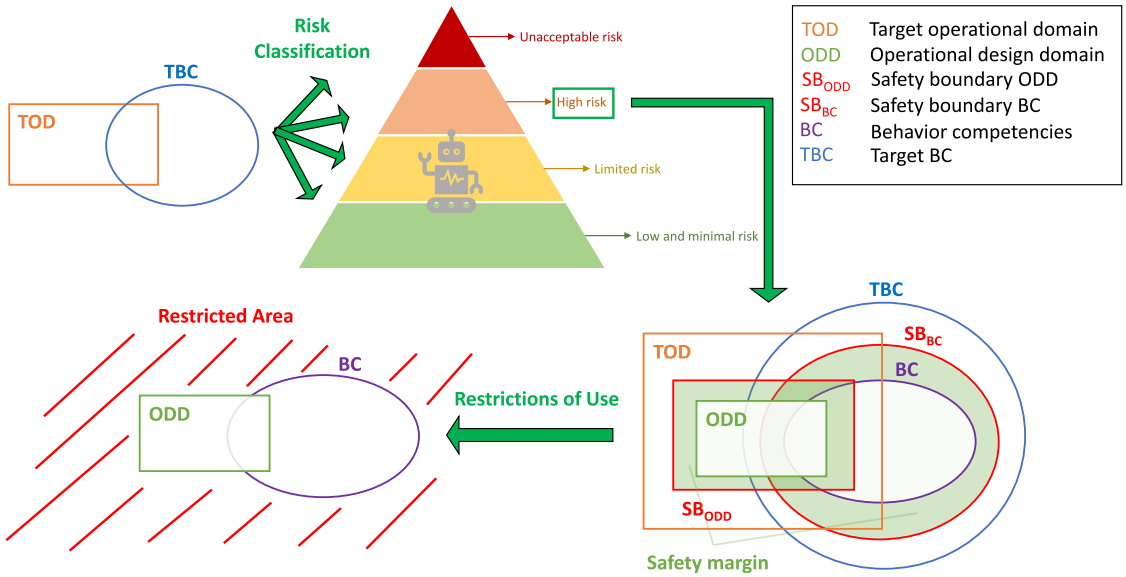

FIGURE 20. Restrictions of use for high-risk AIS.

The initial step involves performing a traceable and explainable risk classification based on the ODD and its associated BC, aligning with the risk categories outlined in the AI Act. The subsequent development of the AIS is contingent on the classified risk, ensuring trustworthiness in accordance with the proposed seven trustworthiness requirements. The defined ODD and BC limits, coupled with an implemented safety margin, describes the covered ODD and BC. This coverage serves as the foundation for formulating restrictions on use. The concept of ODD and its associated BC facilitates the evaluation of various residual risks, serving as qualitative measures to specify acceptable restrictions of use for high-risk AIS. Utilizing the covered and approved ODD and BC, restrictions of use can be articulated using ODD and BC elements for a unique and standardized description. The covered ODD and BC enable the definition of a structured "restricted area," establishing clear boundaries for intended deployment and explicitly delineating what is not intended during the deployment phase. This approach ensures a systematic and transparent formulation of restrictions on use for high-risk AIS.

3) CONCLUSION

The introduced concept of ODD and its associated BC not only enables a traceable and explainable risk classification process aligning with the specified risk categories in the AI Act but also facilitates the formulation of usage restrictions based on specific ODD and BC elements. From a user acceptance perspective, clear and unambiguous usage restrictions are anticipated to be crucial in instilling confidence in general-purpose AIS. Ensuring transparency in usage restrictions minimizes the likelihood of misinterpretations, contributing to greater user trust. From a technological standpoint, a predictable risk classification of the AIS during development is expected to enhance innovation. Companies can chart a clear development path from the outset, avoiding negative surprises at the end of the development cycle when deployment activities are imminent. This predictability fosters a conducive environment for innovation and promotes responsible and transparent development practices in the AI landscape.

B. SUB-GOAL 2: TRUSTWORTHY ASSURANCE/ ARGUMENTATION

1) MOTIVATION

The second sub-goal in achieving trustworthy and human-centric AIS is establishing trustworthy assurance argumentation. Following the identification of potential restrictions of use (or their absence) in the first sub-goal, the objective here is to provide a meaningful argumentation for considering a specific AIS as trustworthy. The key to achieving this lies in the high-level trustworthy assurance process introduced in Figure 17. Specifically, this process should ensure the trustworthiness of the AIS by employing relevant metrics (see section IV-E).

2) METHOD

Constructing an argument for trustworthiness, akin to arguing for safety or any other goal, necessitates the collection of meaningful evidence that can be utilized in a comprehensive argument for achieving the desired goal—in this case, trustworthiness for an AIS. A potential approach for gathering the necessary evidence is to design and deploy a dedicated process: the trustworthy assurance process. By executing the proposed process using the defined boundaries of ODD and BC, evidence generation for trustworthiness is facilitated. The specific argument for trustworthiness may vary based on the use case. In general, certain sub-arguments can

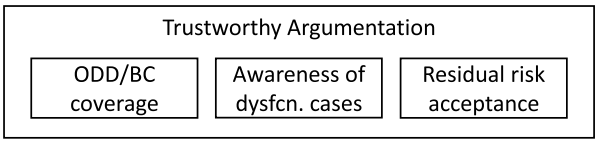

FIGURE 21. The trustworthy argumentation as part of the overall trustworthy assurance process.

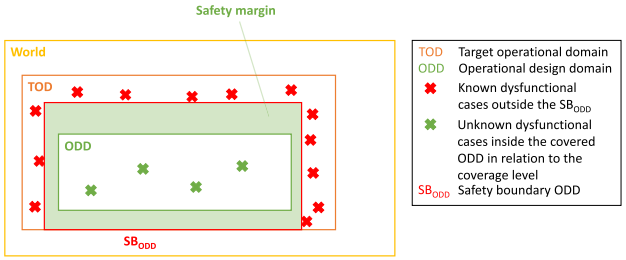

FIGURE 22. Potential dysfunctional cases across the different types of operational domains.

be generated concurrently, either needing to be achieved in combination or assessed independently (see Figure 21). The first sub-argument is ODD/BC coverage, grounded in the core concept that defining the ODD and BC for the AIS establishes the potential input space and its respective boundary. Consequently, this allows the calculation of coverage, essentially representing a fraction of tested versus potential scenarios (parameter combinations). This concept is further detailed for the ADS domain in [81]. The second sub-argument relates to the awareness of dysfunctional cases, as explained in section IV-C. The third sub-argument involves residual risk acceptance. Leveraging the introduced concepts of residual risk, the assurance process can be employed to determine the pertinent evidence for evaluating the applicable residual risk, serving as a basis for informed decision-making.

3) CONCLUSION

Employing a process to determine the trustworthiness of an AIS brings about several significant advantages. Firstly, it facilitates deterministic decision-making grounded in the generated evidence and defined validation targets. This underscores the importance of establishing meaningful metrics for trustworthiness across all sub-arguments and emphasizes the necessity for discussions to reach consensus on widely accepted validation targets. Secondly, the quantitative approach not only provides a basis for informed decision-making but also opens up possibilities for further optimization. It enables the determination of potential assurance efforts in advance and allows for structured adaptations of the AIS itself or its boundaries, such as the ODD and BC.

C. SUB-GOAL 3: AWARENESS OF DYSFUNCTIONAL CASES

1) MOTIVATION

Recent studies indicate that many deployed AI products, particularly in the past decade, exhibit some form of dysfunctionality [96]. While this may be inconsequential for products where AI serves as an additional benefit and is not relied upon to ensure the product's functionality, the landscape has evolved with the introduction of LLMs and other AIS'. These systems often perform critical tasks, introducing safety risks, especially in high-risk applications [97]. Dysfunctional cases capture instances where the AIS fails to fulfill its intended functionality. The underlying causes can be categorized using taxonomies such as the one in [96] and [98], encompassing reasons ranging from conceptual impossibilities to engineering failures and post-deployment issues. Irrespective of the cause, the ultimate effect is the dysfunctionality of the AIS, which is particularly undesired in high-risk applications. As future AIS are held accountable for warranties, fraud, or product liability, awareness of dysfunctionalities becomes crucial. Hence, the awareness of dysfunctional cases is introduced as a sub-argument in the trustworthy argumentation (section IV-B).

2) METHOD

The core idea is to leverage the concepts of ODD and BC, introduced for AIS, to generate evidence for this sub-argument using the trustworthy assurance process (see Figure 17). As explained in the first sub-goal, the metric of ODD/BC coverage is used for a general decision on AIS deployment. However, even in target ODDs where the respective BCs are adequately covered, dysfunctional cases can still occur. The occurrence of such dysfunctionalities directly correlates with the number of tests, although the exact distribution of such cases is unknown. Figure 22 qualitatively displays this. The ODD/BC coverage metric focuses on the overall risk of deployment across the complete ODD and BC range, determining the final target ODD for deployment without making concrete statements about the potential of dysfunctional cases in the covered sections. However, dysfunctional cases could be distributed unevenly across the ODD/BC, leading to the reduction of ODD boundaries in certain parts if individual residual risks exceed a predefined threshold. Consequently, the awareness of dysfunctional cases within the target ODD can be argued using the ODD coverage approach used for the sub-goal in section IV-B, enhanced with the capability to statistically determine the probability of a dysfunctional case.

3) CONCLUSION

Increasing the number of test cases reduces overall variance, thereby decreasing uncertainty about potential critical parameter combinations leading to dysfunctional cases and subsequently lowering the overall residual risk. The metric of target ODD coverage evaluation allows for a statement regarding the relationship between coverage and variance across the complete target ODD. However, the residual risk for dysfunctional cases is not uniformly distributed across

the target ODD, and it may vary for each BC. To enhance awareness of such cases and describe their occurrence, a detailed investigation becomes necessary. The outcome of such an analysis could be a more refined safety margin that ensures specific boundaries regarding the residual risk for dysfunctional cases across the target ODD. While a potential method for determining this, based on the ODD coverage metric, has been introduced, further investigation is required to validate its effectiveness.

D. SUB-GOAL 4: SCENARIO DATABASES AND DATASETS

1) MOTIVATION

In order to implement safety argumentation approaches based on the ODD and BC coverage approach, awareness of dysfunctional cases and associated residual risks, scenarios, and data sets [99], [100] used within the pre-deployment phase of the AI lifecycle must be labelled with ODD and BC elements. Without these labels, no meaningful evaluation within the trustworthy assurance process is possible, as the link to the target ODD and BC is missing.

2) METHOD

Scenarios and datasets serve as the foundation for building trustworthy AIS in terms of law, ethics, and robustness, as depicted in Figure 23. ODD and BC labels integrated into scenarios and datasets facilitate a well-prepared pre-deployment phase for AIS, ensuring readiness for verification, validation, and deployment, and forming a robust trustworthiness argument. According to the introduced trustworthiness assurance process, ODD and BC labels must be established when scenarios and datasets enter the process. Additionally, the intelligent separation of training and validation data, following principles like the k-fold approach [101], becomes a crucial aspect within the trustworthy assurance process of AIS. This necessitates adopting a new data management paradigm, emphasizing the consideration of which parts of the dataset and scenarios will be utilized for design, development, verification, validation, and ultimately, the deployment/certification of the AIS. The outlined methodology is applicable to both synthetic and real datasets and scenarios. Data aquisition is out of scope of this article, however it can be considered as required input of the overall trustworthy assurance process as it is a necessary part of the AIS lifecycle.

3) CONCLUSION

Labelled ODD and BC elements within scenarios and databases are a prerequisite for constructing a robust trustworthiness argumentation, relying on ODD and BC coverage, awareness of dysfunctional cases, and associated residual risks. Moreover, adopting a data management paradigm that strategically determines which parts of the dataset and scenarios will be employed throughout the design, development, verification, validation, and ultimately the deployment/certification of the AIS is a crucial element for ensuring a trustworthy AIS.

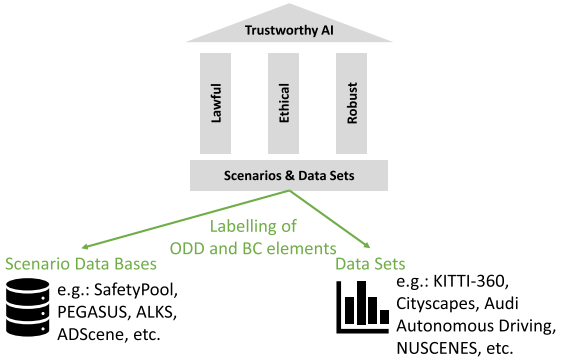

FIGURE 23. Labelling of ODD and BC elements within scenarios and data sets. All six scenario layers include ODD and BC elements to be labelled adequately.

E. SUB-GOAL 5: METRICS AND KPIS

1) MOTIVATION

Defining metrics and KPIs is crucial for ensuring the trustworthiness of AIS. These metrics are integral to the trustworthy assurance process, serving as the basis for evidence in the overall trustworthiness argumentation. Additionally, metrics and KPIs play a significant role throughout the AI lifecycle (section IV-F (sub-goal 6)), influencing decisions such as deployment and monitoring performance in the post-deployment phase.

2) METHOD

The absence of defined metrics and KPIs hinders the quantification of trustworthiness requirements and the establishment of target thresholds. In the pre-deployment phase, leading metrics guide decision-making based on assumptions, as real-world deployment has not yet occurred. On the other hand, lagging metrics, relying on data from actual deployment, are employed in the post-deployment phase to evaluate AIS performance and inform decisions about necessary adaptations. The distinction between leading and lagging metrics is essential, with leading metrics playing a crucial role in the pre-deployment phase and lagging metrics coming into play in the post-deployment phase. The dynamic nature of the AI lifecycle, as illustrated in Figure 11, emphasizes the ongoing need for metrics and KPIs to ensure that residual risks remain below acceptable thresholds. It is noteworthy that even in repeated AI lifecycle iterations, leading metrics continue to be pivotal for deployment decisions, particularly when adaptations in the AIS may disqualify previously gathered data and evaluated lagging metrics. As leading and lagging metrics operate on different types of data, the argumentation strategies for trustworthiness need to be tailored accordingly.

3) CONCLUSION

Many metrics and KPIs are required to support the respective argumentation strategies for each individual trustworthiness

requirement. However, not only the metrics and KPIs itself, but also the respective threshold are a future subject for harmonization and require increased focus across all relevant stakeholders. Only if these metrics, KPIs and respective thresholds are explainable to a wider audience, including the public domain, the acceptance of the overall trustworthy assurance process can be guaranteed. Another important aspect is the distinction between the pre- and post-deployment phase when deciding for metrics and KPIs, due to the different available data sources.

F. SUB-GOAL 6: AI PRODUCT LIFE CYCLE

1) MOTIVATION

The concept of ODD and its related BC is essential for the entire trustworthy and human centric trustworthiness assurance assessment. It starts from the very beginning with the specification of requirements based on ODD and BC attributes, through the development phase, deployment, operational use, including monitoring, to the evaluation and analysis phase. This strong link underlines the importance of preparing the whole AI lifecycle to be ODD and BC element driven.

2) METHOD

The introduction of the ODD and BC concept at all stages within the AI lifecycle enables a consistent and scalable development and deployment strategy of AIS. As a result, a traceable and consistent evaluation of the proposed residual risks along the AI lifecycle is possible which is essential to form an adequate deployment argumentation, see Figure 17. In that sense, the proposed trustworthiness assurance process is fully embedded in the AI lifecycle in the pre- and post-deployment phase (see Figure 24) which enables a continuous evaluation of the restrictions of use, being prepared for continuous updates with the lifetime of AIS e.g., caused by software updates, ODD and/or BC extensions, a structured reporting of dysfunctional cases etc. This also enables a structured certification and re-certification task of AIS properly.

3) CONCLUSION

Without having enrolled the ODD and BC concept within the entire AI lifecycle addressing all individual sub-goals as well, it is not possible to make use of all essential benefits targeting a consistent trustworthiness assurance assessment for AIS. In that sense, all sub-goals are fully connected to the AI lifecycle, feeding in essential elements to enable the overall goal of developing and deploying trustworthy and human centric AIS compliant to the AI Act.

G. SUB-GOAL 7: HUMAN FACTORS

1) MOTIVATION

Humans, by their nature, possess a strong desire to control and understand the world around them. When they perceive a potential loss of control and awareness regarding ongoing activities, they instinctively experience fear, leading to diminished confidence and trust. This inclination extends to complex systems operating in conjunction with humans, such as AIS. Presently, it is evident that a clear and easily understandable description of the operational scope of AIS is lacking. A notable instance is the introduction of automation levels for ADS, which proved too ambiguous to delineate the targeted operational domain and corresponding boundaries. This ambiguity sparked extensive discussions not only in the consumer sector but also within the industry responsible for developing these systems. Consequently, there was a lack of a clear understanding of the operating conditions, such as Level-3 ADS, as depicted in the left part of Figure 25, contributing to a reduction in trust in emerging technologies like ADS.

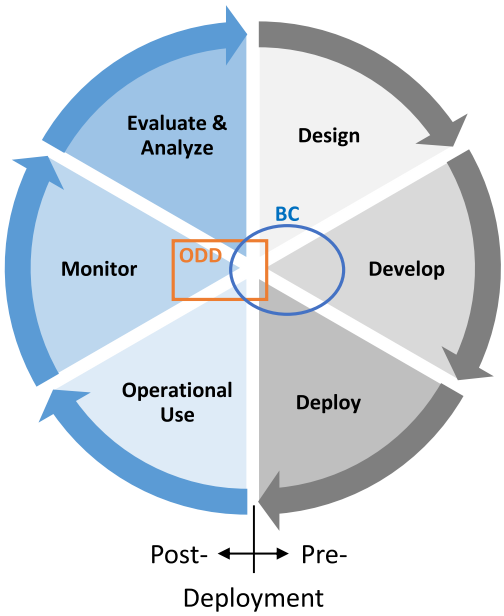

FIGURE 24. ODD and BC driven AI lifecycle.

2) METHOD

Humans find comfort in knowing the exact capabilities and limitations of an AIS operating in close proximity to them. To articulate these capabilities and associated limitations/boundaries, the concepts of ODD and BC emerge as highly promising candidates, as illustrated in the right part of Figure 25. As outlined in Sub-Goal 1 (see section IV-A), the concepts of ODD and BC can rationalize the corresponding risk level and the resulting usage restrictions of any AIS, marking a significant stride in regaining confidence and trust in AIS. Another aspect tied to human factors involves employing meaningful and easily understandable symbols to represent specific ODD and BC elements in a standardized manner. This ensures that communication between AIS and humans remains rapid, clear, and simple, without the necessity of reading detailed texts describing ODD and BC elements. Furthermore, the trustworthiness assurance

assessment process must be traceable and explainable, at least within the AIS value chain deploying such systems. This traceability enhances trustworthiness in AIS by providing clarity in the assessment process.

3) CONCLUSION

Human factors represent a very important pillar within the AI Act to realize trustworthy and human-centric AIS. Following that approach, simple, clear and traceable descriptions of the operating condition and its limitations are essential to build confidence and trust in AIS. Addressing this issue, the concept of ODD and BC supports the human desire to control and understand their surroundings.

V. THE ROLE OF STANDARDS

A. INTRODUCTION

1) WHY ARE STANDARDS IMPORTANT?

AIS classified as high risk must adhere to AI trustworthiness requirements. Legal requirements, often articulated as essential provisions, are outlined in high-level terms, as exemplified by the AI Act (refer to section I-B). Notably, the AI Act does not prescribe the technical methods for meeting these requirements. Instead, in alignment with European legislation under the New Legislative Framework, it establishes fundamental high-level requirements safeguarding public interests. Additionally, it mandates the creation of European harmonized standards essential for products to align with these requirements [102], [103]. Harmonized standards will play a pivotal role in defining technical solutions to fulfill these requirements. However, economic operators have the flexibility to employ technical solutions other than harmonized standards to demonstrate compliance. Harmonized standards thus serve as a crucial tool in implementing the legislation, contributing to the specific objective of ensuring the safety and trustworthiness of AIS. Upon the AI Act coming into force, it will be supported by a set of technical specifications developed by European Standardization Organizations (ESOs). The primary international and European standards development organizations (SDOs) involved in this process include ISO/IEC, ETSI, IEEE, and ITU-T. Importantly, the development of AI standards and technical specifications in support of the AI Act does not commence from scratch. ESOs have the capability to leverage existing standards and technical specifications, facilitated by cooperation agreements. This approach, adopting existing international work, proves to be the most efficient way to prevent duplication of effort and significantly reduce the time required for the development of the diverse range of standards mandated by the forthcoming AI regulation.

The following sources are essential for the collection of existing standards [102], [103]:

1) Surveys on AI standardization, e.g., the CEN/CELEC Focus Group on AI White Paper, the final report of the H2020 StandICT.eu project, the technical report on ''Standards for AI Governance'', etc.
2) Scientific publications, e.g., Journal of ICT Standardisation manuscripts
3) Content from ESOs and SDOs, e.g., ETSI, IEEE, CEN/CENELEC, ISO/IEC JTC1, ITU-T
4) AI standardization roadmaps, e.g., ETSI, CEN/CENELEC, ISO/IEC JTC1, the German AI standardization roadmap, ITU-T AI
5) Focus groups, committees, and projects working on AI standardization, e.g., ISO/IEC JTC1-SC2, EC - CEN CENELEC Focus Group on Artificial Intelligence, the Expert Advisory Group (EAG) of the new StandICT.eu project, the EU-Japan AI Joint Committee, etc.
6) Specific events dealing with ICT and AI standardization, e.g., DG CNECT webinars, JRC workshops, JRC PolicyLab, DG GROW-CNECT-JRC meetings, etc.

2) STANDARDS DEPENDENCIES

Standard specifications usually build on other existing standards to ensure coherence and to avoid conflict and duplication. Typically, the development of a new standard builds on one or more underpinning standards. The underpinning standards may in turn be linked to one or more foundational standards. Therefore, when an AI system developer selects a first-level implementation standard, he/she generally discovers one or more propaedeutic standards that he/she must comply with, the second-level standards [102], see Figure 26.

B. ISO/IEC STANDARDIZATION LANDSCAPE

Many relevant ISO/IEC standards are already published or in the pipeline. Table 1 summarises the identified ISO/IEC standards linked to the stated requirements of the AI Act, enabling trustworthy and human-centred AIS [102].

Operationalization indicators allowed significant gaps to be identified at the level of certain sub-requirements of the AI Act: Data and Data Governance, Technical Documentation and Risk System Management. A set of twelve essential operationalization and suitability standards relevant to the eight AI Act requirements was identified, see Figure 27. From these twelve standards, a core group of six standards were identified that are considered highly relevant [102].

C. IEEE STANDARDIZATION LANDSCAPE

The analysis presented in [103] systematically examines a set of 8 IEEE standards. This set encompasses standards from the IEEE 7000 series, specifically designed for ethical autonomous and intelligent systems. It also includes chosen suites of certification criteria from the IEEE Ethics Certification Program for Autonomous and Intelligent Systems (ECPAIS). The documents within the 7000 series specifically address concerns emerging at the crossroads of technology and ethics, with a pronounced emphasis on Autonomous and Intelligent Systems (A/IS). Consequently, these standards serve as highly relevant references within the context of the human-centered perspective embedded in the proposed European AI Regulation, particularly in relation to risks to fundamental rights. Distinguishing itself from

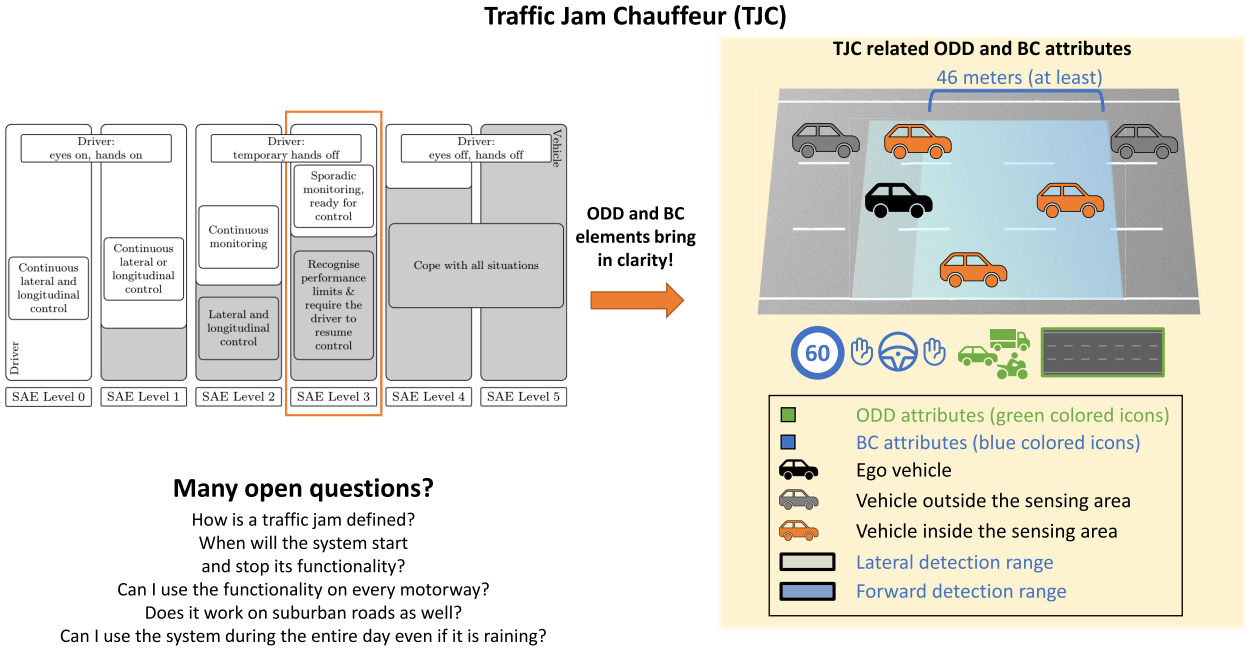

FIGURE 25. Traffic Jam Chauffeur specified via ISO 22736:2021 [36] levels of automation (left part) vs. ODD and its related BC attributes (right part).

TABLE 1. ISO/IEC standardization landscape [102].

Requirement	Relevant operationalization standards
Data and data governance	ISO/IEC TS 4213, ISO/IEC 5259-2, ISO/IEC 5259-3, ISO/IEC 5259-4, ISO/IEC 5338, ISO/IEC 5469, ISO/IEC 23894.2, ISO/IEC 24027, ISO/IEC 24029-1, ISO/IEC 24668, ISO/IEC 38507, ISO/IEC 42001, ETSI SAI 002, ETSI SAI 005
Technical documentation	ISO/IEC 23894.2, ISO/IEC 24027, ISO/IEC 42001
Record keeping	ISO/IEC 23894.2
Transparency and information to users	ISO/IEC 23894.2, ISO/IEC 24027, ISO/IEC 24028, ISO/IEC 38507, ISO/IEC 42001
Human oversight	ISO/IEC 23894.2, ISO/IEC 38507, ISO/IEC 42001
Accuracy robustness and cybersecurity	ISO/IEC TS 4213, ISO/IEC 5338, ISO/IEC 5469, ISO/IEC 23894.2, ISO/IEC 24029-1, ISO/IEC 24668, ISO/IEC 42001, ETSI SAI 002, ETSI SAI 003, ETSI SAI 005, ETSI SAI 006
Risk management system	ISO/IEC 5338, ISO/IEC 5469, ISO/IEC 23894.2, ISO/IEC 38507, ISO/IEC 42001
Quality management system	ISO/IEC 5259-3, ISO/IEC 5259-4, ISO/IEC 5338, ISO/IEC 23894.2, ISO/IEC 24029-1, ISO/IEC 38507, ISO/IEC 42001

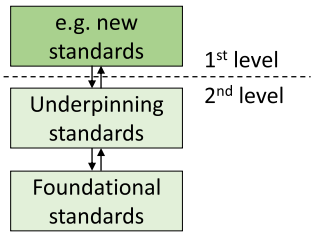

FIGURE 26. Standards dependencies.

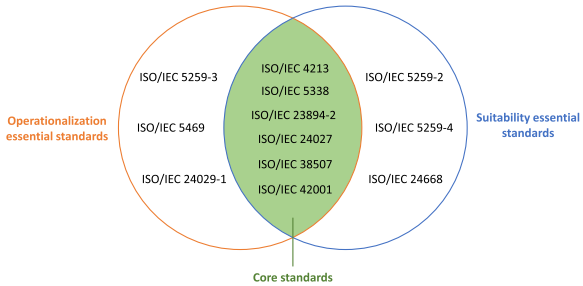

FIGURE 27. Relationship between the operational and suitability essential standards and its related core group [102].

process-oriented standards, the ECPAIS certification suites represent a distinct category of deliverables. These suites complement standards by adopting an outcome-based approach, offering criteria to assess key facets of trustworthy AI. These include accountability, transparency, and the reduction of algorithmic bias. Furthermore, they facilitate the certification of A/IS products, systems, and services based on these criteria. A selected subset of documents from these families is outlined in Table 2. The main findings in terms of observations and recommendations can be summarized as follows [103]:

1) The IEEE 7000 Standard Model Process for Addressing Ethical Concerns during System Design is a

useful reference for operationalizing risk management requirements in AI law. It is worth highlighting its level of descriptiveness and product orientation, as it provides a process for systematically considering and addressing ethical values and risks in the design of an AI system, and translating them into traceable product requirements.

2) The IEEE P7001 Draft Standard for Transparency of Autonomous Systems provides relevant coverage of human oversight aspects. The IEEE P7001 draft standard also provides valuable coverage of record-keeping requirements in the proposed European AI regulation.
3) The IEEE P7003 Draft Standard for Algorithmic Bias Considerations should be considered as a relevant source of technical specifications for operationalizing the AI Act's bias requirements.

The examination of these standards has resulted in the identification of valuable content that can be instrumental in operationalizing requirements related to AI bias, human oversight, record-keeping, and risk management. Simultaneously, the scrutiny of certification criteria suggests that, in the future, these criteria could serve as a foundation for developing implementable methods to verify compliance with the AI Regulation. This ongoing analysis will be expanded in forthcoming reports to encompass additional documents, either from the same series or from other pertinent families of standards, as noted in [103]. Within the IEEE framework, this extension includes selected documents from the 2800 series, focusing on AI governance and licensing issues, specific technologies like deep learning or federated learning, and even concrete application areas of interest in the context of the AI Act, such as healthcare or robotics.

D. ADS SAFETY STANDARDIZATION LANDSCAPE

In product development across the automotive domain, the ISO 26262 [57] standard ensures that automotive systems do not cause hazards due to technical failures from faults in the systems' software or hardware. The ISO/PAS 21448 [60] standard, also known as Safety of the Intended Functionality (SOTIF), ensures that the systems' intended functionality is safe and does not cause hazards, even without technical failures. For AD, there are many factors that can trigger unknown and hazardous scenarios. The safety assessment of AD refers to the SOTIF aspect. It also includes the assessment of the OEDR capabilities of an ADS for the whole ODD.

E. ODD/BC STANDARDIZATION LANDSCAPE

Originating in the ADS safety domain, key standards pertaining to the concept of ODD include three fundamental standards: BSI PAS 1883:2020 [34], SAE J3016 (2021) [104] along with its ISO counterpart ISO 34503:2023 [33]. Additionally, there are user-specific standards such as ASAM OpenODD [105], which focuses on providing a format capable of representing a defined ODD for Connected Automated Vehicles (CAV). Another noteworthy standard is the AVSC Best Practice for describing an ODD:

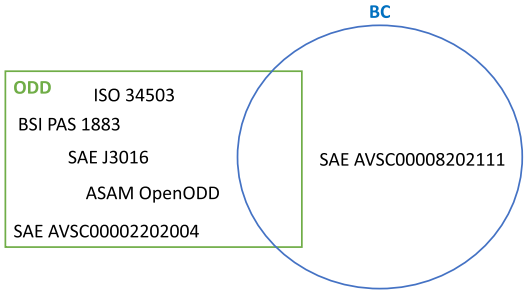

FIGURE 28. ODD/BC standardization landscape.

Conceptual Framework and Lexicon (AVSC00002202004) [106], published by the Automated Vehicle Safety Consortium™(AVSC), which is an industry program of the SAE Industry Technologies Consortia (SAE ITC®) working to rapidly publish best practices that inform and lead to industry-wide standards, promoting the safe deployment of automated driving systems (ADS). Figure 28 illustrates the available ODD-related standards in relation to their BC counterparts. Based on these standards, the development of safety measures rooted in ODD principles has progressed significantly, particularly by formulating compelling deployment arguments. In contrast to the ODD standardization landscape, the situation for BC's is notably different. The sole available standardization-related document is published by the AVSC. The document, titled Best Practice for Evaluation of Behavioral Competencies for Automated Driving System Dedicated Vehicles (AVSC00008202111) [39], focuses on evaluation concepts for BC rather than specifying them (see Figure 28). Harmonized standards for BC are urgently needed to provide a robust argumentation for the trustworthiness of ADS and AIS, aligning with the rationale established in the ODD standardization landscape.

VI. KEY TAKEAWAYS AND CALL FOR ACTION

A. KEY TAKEAWAYS

The assessment of trustworthiness of AIS' requires a comprehensive approach that covers multiple objectives. In the following, the key takeaways gathered from the contributions of this article are summarized:

- In general, the introduction of the ODD and BC concept is relevant for the achievement of all necessary subgoals for achieving trustworthy and human centric AIS'. Also, the consistent introduction and consequent appliance is key.
- The boundaries of AIS' in terms of its application area and capabilities, using the ODD and BC concepts must be done in a way that is acceptable to both engineers and end users.
- From a technological point of view, a predictable risk classification of the AIS to be developed by companies will boost their innovation power as development uncertainties are kept to a minimum.

TABLE 2. Overview of the analyzed IEEE standards and its related links to the AI Act requirements.

Requirement	Document	Type
Risk Management, Traceability	IEEE 7000 Standard Model Process for Addressing Ethical Concerns during System Design	Standard
Transparency, Human Oversight	IEEE P7001/D4 Draft Standard for Transparency of Autonomous Systems	Standard
Avoidance of Unfair Bias	IEEE P7003/D1 Draft Standard for Algorithmic Bias Considerations	Standard
Social and Environment Well-being	IEEE 7010 Recommended Practice for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-Being	Standard
Data Governance	IEEE P2841 - Framework and Process for Deep Learning Evaluation	Standard
Accountability	IEEE ECPAIS: Accountability Certification Requirements	Certification Criteria
Transparency	IEEE ECPAIS: Transparency Certification Requirements	Certification Criteria
Avoidance of Unfair Bias	IEEE ECPAIS: Bias Certification Requirements	Certification Criteria

- A harmonized process for AIS trustworthiness assurance is essential to enable a standardized and generally accepted deployment procedure across the EU. Overall, such a process needs to generate the required evidence to argue for AIS trustworthiness in a scalable manner.
- The awareness of dysfunctional cases is only achievable using a structured approach that incorporates the respective metrics that allow the quantification of the residual risk. In addition, the reporting of such cases needs to be carried out in a harmonized and understandable manner so that future developments can profit from past failures - similar to best practice approaches in the aviation industry.
- The consistent labelling of the ODD and BC elements within scenes, scenarios as well as databases are a prerequisite for consistent usage of these concepts across the entire AI lifecycle. Without having enrolled the ODD and BC concept within the entire AI lifecycle addressing all individual sub-goals, it is impossible to achieve all essential benefits of a consistent trustworthiness assurance assessment for AIS.
- KPIs and metrics eventually decide about trustworthiness and are therefore key elements with high priority and impact. Only if these metrics, KPIs and respective thresholds are explainable to a wider audience, including the public domain, the acceptance of the overall trustworthy assurance process can be guaranteed.
- Human factors represent a very important pillar within the AI Act to realize trustworthy and human-centric AIS. Hence, simple, clear and traceable descriptions of the operating condition and its limitations are essential to build confidence and trust in AIS.

To conclude: the concept of ODD and the related BC support the two main objectives of the AI Act, namely, to enable the deployment of a human-centric and trustworthy AIS in Europe.

B. CALL FOR ACTIONS

Based on the outlined key takeaways collected during the Trustworthiness Assurance Assessment, the following Calls for Action have been crystallized to address the current white spots and gaps and to take a giant step towards the realization of a trustworthy and human-centric AIS. The following calls for action are prioritized from top to bottom:

1) **Call for Deployment:** Deploy the introduced ODD and associated BC to the AI domain along the entire AIS value chain and AI lifecycle, including communication and training activities to get all AIS value chain partners on board. Only when all partners follow the same approach can all the associated benefits be realized efficiently and effectively. This vision is based on a common understanding along all partners, not forgetting the human factors' perspective.
2) **Call for Collaboration:** Without the implementation of the ODD and BC concept along the whole AIS value chain, it is impossible to learn from mistakes and to refine the concept based on the experience gained in different AIS application areas. This means that the very important task of refining, extending and optimizing the proposed concept depends heavily on a strong collaborative attitude along the AIS value chain. Learning together is one of the main challenges for the future. Besides consensus building for trustworthiness, argumentation is seen as a highly relevant collaborative task.
3) **Call for Standardization:** Based on the previous two calls for deployment and collaboration, the need for standardization is the next logical consequence, addressing several issues. Firstly, standardization activities in the field of BC need to be initiated in the short term as this is the most urgent call. Secondly, a further refinement of the ODD standard based on the experience gained from AIS applications is highly relevant to have a solid standard applicable to numerous domains related to AIS. In particular, the lessons learned from various use cases need to be fed back to the standardization bodies for consideration in subsequent updates of the relevant standards. Thirdly, the definition of standardized KPIs and metrics is essential to assess the trustworthiness of AIS, as no evidence can be produced without harmonized and accepted KPIs and metrics. Finally, ready-to-use ODD

and BC standards enable standardized, labelled data sets and scenarios that can be used consistently by all partners in the AIS value chain.

4) **Call for Sandboxing:** A neutral playground in the form of a regulatory sandbox, which is a controlled environment that facilitates the development, testing, and validation of innovative AI before it is brought to market, is essential to overcome companies' fears and uncertainties about whether they are compliant with the AI Act. The sandbox will also allow participants to use personal data to promote AI innovation without prejudice to the requirements of the GDPR. This will boost the innovation potential of companies in general and prevent the loss of promising approaches in Europe due to regulatory fears.

To conclude: the calls for action represent key challenges to be addressed by the entire AIS value chain, which are of significant importance to take the next step in realizing trustworthy and human-centric AIS together. In that sense, it is very important to note that all calls for actions can only be addressed properly together, not by a single partner.

REFERENCES

[1] *Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts*. Accessed: Nov. 9, 2023. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206

[2] European Union. (2012). *Charter of Fundamental Rights of the European Union*. [Online]. Available: https://www.refworld.org/docid/3ae6b3b70.html

[3] Y. Cao, Q. Z. Sheng, J. McAuley, and L. Yao, "Reinforcement learning for generative AI: A survey," 2023, *arXiv:2308.14328*.

[4] A. Garg. *What is ChatGPT, and Its Possible Use Cases?* Insights—Web Mobile Development Services Solutions. Accessed: Aug. 24, 2023. [Online]. Available: https://www.netsolutions.com/insights/what-is-chatgpt/

[5] S. Mohamadi, G. Mujtaba, N. Le, G. Doretto, and D. A. Adjeroh, "ChatGPT in the age of generative AI and large language models: A concise survey," 2023, *arXiv:2307.04251*.

[6] F. Martánez-Plumed, F. Caballero, D. Castellano-Falcón, D. Fernández-Llorca, E. Gómez, I. Hupont-Torres, L. Merino, C. Monserrat, and J. Hernández-Orallo, "AI watch, revisiting technology readiness levels for relevant artificial intelligence technologies," Joint Res. Centre (Eur. Commission), Publications Office Eur. Union, Seville, Spain, Tech. Rep. JRC129399, 2022, doi: 10.2760/495140.

[7] Gartner. (2023). *What's New in Artificial Intelligence From the 2023 Gartner Hype CycleT*. [Online]. Available: https://www.gartner.com/en/articles/what-s-new-in-artificial-intelligence-from-the-2023-gartner-hype-cycle

[8] M. Tambiama, "General-purpose artificial intelligence," Eur. Parliament, Brussels, Belgium, Tech. Rep. PE 745.708, 2023.

[9] S. Burton and R. Hawkins. *Assuring the Safety of Highly Automated Driving: State-of-the-art and Research Perspectives*. Accessed: Sep. 10, 2023. [Online]. Available: https://www.york.ac.uk/media/assuring-autonomy/publications/Assuring%20Autonomy%20International%20Programme%20-%20Safety%20Assurance%20of%20Highly%20Automated%20Driving.pdf

[10] High-Level Expert Group on Artificial Intelligence. *A Definition of Artificial Intelligence: Main Capabilities and Scientific Disciplines*. Accessed: Jun. 6, 2023. [Online]. Available: https://digital-strategy.ec.europa.eu/en/library/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines

[11] European Commission. *Artificial Intelligence for Europe*. Accessed: Aug. 10, 2023. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52018DC0237

[12] D. F. Llorca and E. G. Gutierrez, "Artificial intelligence in autonomous vehicles: Towards trustworthy systems," Eur. Commission, Brussels, Belgium, Tech. Rep. JRC128170, 2022.

[13] S. Samoili, M. López Cobo, E. Gómez, G. De Prato, F. Martínez-Plumed, and B. Delipetrev, "AI watch: Defining artificial intelligence: Towards an operational definition and taxonomy of artificial intelligence," Publications Office Eur. Union, Seville, Spain, Tech. Rep. JRC118163, 2020, doi: 10.2760/382730.

[14] S. Samoili, M. López Cobo, B. Delipetrev, F. Martinez-Plumed, E. Gómez, and G. De Prato, "AI watch, defining artificial intelligence 2.0: Towards an operational definition and taxonomy for the AI landscape," Joint Res. Centre (Eur. Commission), Publications Office Eur. Union, Seville, Spain, Tech. Rep. JRC126426, 2021, doi: 10.2760/019901.

[15] European Council. (2023). *Artificial Intelligence Act: Council and Parliament Strike a Deal on the First Rules for Ai in the World*. [Online]. Available: https://www.consilium.europa.eu/en/press/press-releases/2023/12/09/artificial-intelligence-act-council-and-parliament-strike-a-deal-on-the-first-worldwide-rules-for-ai/

[16] *Recommendation of the Council on Artificial Intelligence*, OECD, Paris, France, OECD/LEGAL/0449, 2023.

[17] N. A. Smuha, E. Ahmed-Rengers, A. Harkens, W. Li, J. MacLaren, R. Piselli, and K. Yeung. *How the EU Can Achieve Legally Trustworthy AI: A Response To the European Commission's Proposal for an Artificial Intelligence AC*. Accessed: Aug. 18, 2023. [Online]. Available: https://papers.ssrn.com/abstract=3899991

[18] M. Veale and F. Zuiderveen Borgesius, "Demystifying the draft EU artificial intelligence act—Analysing the good, the bad, and the unclear elements of the proposed approach," *Comput. Law Rev. Int.*, vol. 22, no. 4, pp. 97–112, Aug. 2021. [Online]. Available: https://www.degruyter.com/document/doi/10.9785/cri-2021-220402/html?lang=en

[19] High-Level Expert Group on Artificial Intelligence. *Policy and Investment Recommendations for Trustworthy AI*. Accessed: May 5, 2023. [Online]. Available: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60343

[20] *The Assessment List for Trustworthy Artifical Intelligence (ALTAI) for Self Assessment*, High Level Expert Group Artif. Intell., Brussels, Belgium, 2020.

[21] T. Buocz, S. Pfotenhauer, and I. Eisenberger, "Regulatory sandboxes in the AI Act: Reconciling innovation and safety?" *Law, Innov. Technol.*, vol. 15, no. 2, pp. 357–389, Jul. 2023, doi: 10.1080/17579961.2023.2245678.

[22] (2023). *SANDBOXING the AI ACT—Testing the AI Act Proposal With Europe's Future Unicorns*. [Online]. Available: https://cdn.digitaleurope.org/uploads/2023/06/DIGITAL-EUROPE-SANDBOXING-THE-AI-ACT_FINAL_WEB_SPREADS.pdf

[23] P. Hacker, A. Engel, and M. Mauer, "Regulating ChatGPT and other large generative AI models," in *Proc. ACM Conf. Fairness, Accountability, Transparency*, Jun. 2023, pp. 1112–1123, doi: 10.1145/3593013.3594067.

[24] J. M. Wing, "Trustworthy AI," *Commun. ACM*, vol. 64, no. 10, pp. 64–71, 2021, doi: 10.1145/3448248.

[25] B. Li, P. Qi, B. Liu, S. Di, J. Liu, J. Pei, J. Yi, and B. Zhou, "Trustworthy AI: From principles to practices," *ACM Comput. Surv.*, vol. 55, no. 9, pp. 1–46, Sep. 2023, doi: 10.1145/3555803.

[26] J. Marques-Silva and A. Ignatiev, "Delivering trustworthy AI through formal XAI," in *Proc. AAAI Conf. Artif. Intell.*, Jun. 2022, vol. 36, no. 11, pp. 12342–12350. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/21499

[27] C. González-Gonzalo, E. F. Thee, C. C. W. Klaver, A. Y. Lee, R. O. Schlingemann, A. Tufail, F. Verbraak, and C. I. Sánchez, "Trustworthy AI: Closing the gap between development and integration of AI systems in ophthalmic practice," *Prog. Retinal Eye Res.*, vol. 90, Sep. 2022, Art. no. 101034. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1350946221000951

[28] V. Chamola, V. Hassija, A. R. Sulthana, D. Ghosh, D. Dhingra, and B. Sikdar, "A review of trustworthy and explainable artificial intelligence (XAI)," *IEEE Access*, vol. 11, pp. 78994–79015, 2023. [Online]. Available: https://ieeexplore.ieee.org/document/10188681/citations#citations

[29] D. De Silva and D. Alahakoon, "An artificial intelligence life cycle: From conception to production," *Patterns*, vol. 3, no. 6, Jun. 2022, Art. no. 100489. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666389922000745

[30] M. Poretschkin, A. Schmitz, M. Akila, L. Adilova, D. Becker, A. B. Cremers, D. Hecker, S. Houben, M. Mock, J. Rosenzweig, J. Sicking, E. Schulz, A. Voss, and S. Wrobel, "Guideline for trustworthy artificial intelligence—AI assessment catalog," 2023, arXiv:2307.03681.

[31] S. J. Russell, P. Norvig, M.-w. Chang, J. Devlin, A. Dragan, D. Forsyth, I. Goodfellow, J. Malik, V. Mansinghka, J. Pearl, and M. J. Wooldridge, *Artificial Intelligence: A Modern Approach* (Pearson Series in Artificial Intelligence), 4th ed. London, U.K.: Pearson, 2022.

[32] P. Hamm, M. Klesel, P. Coberger, and H. F. Wittmann, "Explanation matters: An experimental study on explainable AI," *Electron. Markets*, vol. 33, no. 1, p. 17, Dec. 2023, doi: 10.1007/s12525-023-00640-9.

[33] *Road Vehicles—Taxonomy for Operational Design Domain for Automated Driving Systems*, Standard ISO 34503, Int. Org. Standardization, 2021. [Online]. Available: https://www.iso.org/standard/78952.html

[34] *BSI Standards Limited 2020*, Standard PAS 1883, 2020. [Online]. Available: https://www.bsigroup.com/en-GB/CAV/pas-1883/

[35] J. Erz, B. Schutt, T. Braun, H. Guissouma, and E. Sax, "Towards an ontology that reconciles the operational design domain, scenario-based testing, and automated vehicle architectures," in *Proc. IEEE Int. Syst. Conf. (SysCon)*, Mar. 2022, pp. 1–8. [Online]. Available: https://ieeexplore.ieee.org/document/9773840/

[36] *Taxonomy and Definitions for Terms Related To Driving Automation Systems for on-Road Motor Vehicles*, Standard ISO/SAE PAS 22736, 2021. [Online]. Available: https://www.iso.org/standard/73766.html

[37] L. Lili, V. S. Singh, S. Hon, D. Le, K. Tan, D. Zhang, and D. Chai, "AI-based behavioral competency assessment tool to enhance navigational safety," in *Proc. Int. Conf. Electr., Comput., Commun. Mechatronics Eng. (ICECCME)*, Oct. 2021, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/document/9590891/

[38] M. E. Derro and C. R. Williams, "Behavioral competencies of highly regarded systems engineers at NASA," in *Proc. IEEE Aerosp. Conf.*, Mar. 2009, pp. 1–17. [Online]. Available: http://ieeexplore.ieee.org/document/4839712/

[39] *AVSC Best Practice for Evaluation of Behavioral Competencies for Automated Driving System Dedicated Vehicles (ADS-DVs)*, Automated Vehicle Saf. Consortium, PA, USA, 2021.

[40] *Guidelines for Regulatory Requirements and Verifiable Criteria for ADS Safety Validation*, UNECE, Geneva, Switzerland, 2023.

[41] E. Thorn, S. Kimmel, and M. Chaka. (2018). *A Framework for Automated Driving System Testable Cases and Scenarios*. [Online]. Available: https://www.nhtsa.gov/sites/nhtsa.gov/files/documents/13882-automateddrivingsystems_092618_v1a_tag.pdf

[42] S. Khastgir. (2023). *Cross-Domain Safety Assurance for Automated Transport Systems*. [Online]. Available: https://warwick.ac.uk/fac/sci/wmg/research/cav/vandv/ukriflf/2023crossdomainsafety_online_final_v2.0.pdf

[43] M. Hirschle, D. Kirov, R. Aievola, S. Sinisi, S. Iovino, and J. Adamy, "Scenario-based methods for machine learning assurance," in *Proc. IEEE/AIAA 42nd Digital Avionics Syst. Conf. (DASC)*, 2023, pp. 1–10. [Online]. Available: https://ieeexplore.ieee.org/document/10311114

[44] J. Krank. (2020). *Robo-Crop: The Imminence of Autonomous Technology in Agriculture*. [Online]. Available: https://heinonline.org/HOL/Page?handle=hein.journals/dragl25&id=497&div=&collection=

[45] M. Kläs and L. Sembach, "Uncertainty wrappers for data-driven models: Increase the transparency of AI/ML-based models through enrichment with dependable situation-aware uncertainty estimates," in *Computer Safety, Reliability, and Security*, vol. 11699, A. Romanovsky, E. Troubitsyna, I. Gashi, E. Schoitsch, and F. Bitsch, Eds. Berlin, Germany: Springer, 2019, pp. 358–364, doi: 10.1007/978-3-030-26250-1_29.

[46] D. Sculley, G. Holt, D. Golovin, E. Davydov, T. Phillips, D. Ebner, V. Chaudhary, M. Young, J.-F. Crespo, and D. Dennison, "Hidden technical debt in machine learning systems," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28. New York, NY, USA: Curran Associates, 2015, pp. 1–9. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2015/hash/86df7dcfd896fcaf2674f757a2463eba-Abstract.html

[47] B. Xia, Q. Lu, H. Perera, L. Zhu, Z. Xing, Y. Liu, and J. Whittle, "Towards concrete and connected AI risk assessment (C^2AIRA): A systematic mapping study," 2023, *arXiv:2301.11616*.

[48] Gartner. (2023). *Definition of AI TRiSM—Gartner Information Technology Glossary*. [Online]. Available: https://www.gartner.com/en/information-technology/glossary/ai-trism

[49] S. Une Lee, H. Perera, B. Xia, Y. Liu, Q. Lu, L. Zhu, O. Salvado, and J. Whittle, "QB4AIRA: A question bank for AI risk assessment," 2023, *arXiv:2305.09300*.

[50] D. Piorkowski, M. Hind, and J. Richards, "Quantitative AI risk assessments: Opportunities and challenges," 2022, *arXiv:2209.06317*.

[51] Z. Wang, Y. Huang, L. Ma, H. Yokoyama, S. Tokumoto, and K. Munakata, "An exploratory study of AI system risk assessment from the lens of data distribution and uncertainty," 2022, *arXiv:2212.06828*.

[52] Economic Commission for Europe. (2021). *New Assessment/Test Method for Automated Driving (NATM)*. [Online]. Available: https://unece.org/sites/default/files/2021-04/ECE-TRANS-WP29-2021-61e.pdf

[53] *Road Vehicles—Safety and Cybersecurity for Automated Driving Systems—Design, Verification and Validation*, Standard ISO/TR 4804, 2020. [Online]. Available: https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/08/03/80363.html

[54] F. Favarò, L. Fraade-Blanar, S. Schnelle, T. Victor, M. Pena, J. Engstrom, J. Scanlon, K. Kusano, and D. Smith. (2023). *Building a Credible Case for Safety: Waymo's Approach for the Determination of Absence of Unreasonable Risk*. [Online]. Available: https://storage.googleapis.com/waymo-uploads/files/documents/safety/Waymo%20Safety%20Case%20Approach.pdf

[55] F. Favaro, "Exploring the relationship between 'positive risk balance 'and' absence of unreasonable risk,'" 2021, *arXiv:2110.10566*.

[56] S. Schnelle and F. M. Favaro, "ADS standardization landscape: Making sense of its status and of the associated research questions," 2023, *arXiv:2306.17682*.

[57] *Road Vehicles-Functional Safety*, Standard ISO 26262, 2018.

[58] *Systems and Software Quality Requirements and Evaluation (Square)*, Standard ISO 25010, 2011, p. 34. [Online]. Available: https://www.iso.org/standard/70939.html

[59] Underwriters' Laboratories. *Standard for Evaluation of Autonomous Products*, Standard UL 4600, Underwriters Laboratories, 2020. [Online]. Available: https://books.google.at/books?id=7fH3zQEACAAJ

[60] *Road Vehicles—Safety of the Intended Functionality*, Standard ISO/PAS 21448, 2021. [Online]. Available: https://www.iso.org/standard/70939.html

[61] D. Hendrycks, M. Mazeika, and T. Woodside, "An overview of catastrophic AI risks," 2023, *arXiv:2306.12001*.

[62] M. Schwarz, M. Fistler, M. Loehning, and H. Schittenhelm, "Measuring safety: Positive risk balance and conscientious driver," VVM Project, Munich, Germany, Tech. Rep. S1P4, 2022.

[63] T. Dreossi, D. J. Fremont, S. Ghosh, E. Kim, H. Ravanbakhsh, M. Vazquez-Chanlatte, and S. A. Seshia, "VerifAI: A toolkit for the formal design and analysis of artificial intelligence-based systems," in *Computer Aided Verification*, vol. 11561, I. Dillig and S. Tasiran, Eds. Berlin, Germany: Springer, 2019, pp. 432–442, doi: 10.1007/978-3-030-25540-4_25.

[64] D. J. Fremont, T. Dreossi, S. Ghosh, X. Yue, A. L. Sangiovanni-Vincentelli, and S. A. Seshia, "Scenic: A language for scenario specification and scene generation," 2018, *arXiv:1809.09310*.

[65] W. Huang, X. Zhao, G. Jin, and X. Huang, "SAFARI: Versatile and efficient evaluations for robustness of interpretability," 2022, *arXiv:2208.09418*.

[66] H. Khedr and Y. Shoukry, "DeepBern-Nets: Taming the complexity of certifying neural networks using Bernstein polynomial activations and precise bound propagation," 2023, *arXiv:2305.13508*.

[67] S. Feuerriegel, J. Hartmann, C. Janiesch, and P. Zschech, "Generative AI," 2023, *arXiv:2309.07930*.

[68] A. E. Goodloe, "Assuring safety-critical machine learning-enabled systems: Challenges and promise," *Computer*, vol. 56, no. 9, pp. 83–88, Sep. 2023.

[69] M. Jirotka, B. Grimpe, B. Stahl, G. Eden, and M. Hartswood, "Responsible research and innovation in the digital age," *Commun. ACM*, vol. 60, no. 5, pp. 62–68, Apr. 2017.

[70] H. Herrmann, "What's next for responsible artificial intelligence: A way forward through responsible innovation," *Heliyon*, vol. 9, no. 3, Mar. 2023, Art. no. e14379. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2405844023015864

[71] R. Hawkins, C. Picardi, L. Donnell, and M. Ireland, "Creating a safety assurance case for a machine learned satellite-based wildfire detection and alert system," *J. Intell. Robotic Syst.*, vol. 108, no. 3, p. 47, Jul. 2023, doi: 10.1007/s10846-023-01905-3.

[72] A. Onnes, "Monitoring AI systems: A problem analysis, framework and outlook," 2022, *arXiv:2205.02562*.

[73] D. Hendrycks, N. Carlini, J. Schulman, and J. Steinhardt, "Unsolved problems in ML safety," 2021, *arXiv:2109.13916*.

[74] L. Myllyaho, M. Raatikainen, T. Männistö, T. Mikkonen, and J. K. Nurminen, "Systematic literature review of validation methods for AI systems," *J. Syst. Softw.*, vol. 181, Nov. 2021, Art. no. 111050. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0164121221001473

[75] R. Tsopra et al., "A framework for validating AI in precision medicine: Considerations from the European ITFoC consortium," *BMC Med. Informat. Decis. Making*, vol. 21, no. 1, p. 274, Dec. 2021, doi: 10.1186/s12911-021-01634-3.

[76] A. C. Yu, B. Mohajer, and J. Eng, "External validation of deep learning algorithms for radiologic diagnosis: A systematic review," *Radiol., Artif. Intell.*, vol. 4, no. 3, May 2022, Art. no. e210064, doi: 10.1148/ryai.210064.

[77] R. Hamon, H. Junklewitz, J. I. S. Martin, D. F. Llorca, E. G. Gutierrez, A. H. Alcantara, and A. Kriston, "Artificial intelligence in automated driving: An analysis of safety and cybersecurity challenges," Eur. Commission, Brussels, Belgium, Tech. Rep. JRC127189, 2022.

[78] M. Zeller, T. Waschulzik, R. Schmid, and C. Bahlmann, "Towards a safe MLOps process for the continuous development and safety assurance of ML-based systems in the railway domain," 2023, *arXiv:2307.02867*.

[79] S. Burton, "A causal model of safety assurance for machine learning," 2022, *arXiv:2201.05451*.

[80] R. Adler and M. Klaes, "Assurance cases as foundation stone for auditing AI-enabled and autonomous systems: Workshop results and political recommendations for action from the ExamAI project," 2022, *arXiv:2208.08198*.

[81] P. Weissensteiner, G. Stettinger, S. Khastgir, and D. Watzenig, "Operational design domain-driven coverage for the safety argumentation of automated vehicles," *IEEE Access*, vol. 11, pp. 12263–12284, 2023.

[82] P. Matthias Winter, S. Eder, J. Weissenböck, C. Schwald, T. Doms, T. Vogt, S. Hochreiter, and B. Nessler, "Trusted artificial intelligence: Towards certification of machine learning applications," 2021, *arXiv:2103.16910*.

[83] M. Kwiatkowska and X. Zhang, "When to trust AI: Advances and challenges for certification of neural networks," 2023, *arXiv:2309.11196*.

[84] D. Brajovic, N. Renner, V. Philipp Goebels, P. Wagner, B. Fresz, M. Biller, M. Klaeb, J. Kutz, J. Neuhuettler, and M. F. Huber, "Model reporting for certifiable AI: A proposal from merging EU regulation into AI development," 2023, *arXiv:2307.11525*.

[85] N. Scharowski, M. Benk, S. J. Kuhne, L. Wettstein, and F. Brühlmann, "Certification labels for trustworthy AI: Insights from an empirical mixed-method study," 2023, *arXiv:2305.18307*.

[86] S. Pathrudkar, S. Venkataraman, D. Kanade, A. Ajayan, P. Gupta, S. Khatib, V. S. Indla, and S. Mukherjee, "SAFR-AV: Safety analysis of autonomous vehicles using real world data—An end-to-end solution for real world data driven scenario-based testing for pre-certification of AV stacks," 2023, *arXiv:2302.14601*.

[87] M. Gariel, B. Shimanuki, R. Timpe, and E. Wilson, "Framework for certification of AI-based systems," 2023, *arXiv:2302.11049*.

[88] P. Cihon, M. J. Kleinaltenkamp, J. Schuett, and S. D. Baum, "AI certification: Advancing ethical practice by reducing information asymmetries," 2021, *arXiv:2105.10356*.

[89] T. Tommasi, S. Bucci, B. Caputo, and P. Asinari, "Towards fairness certification in artificial intelligence," 2021, *arXiv:2106.02498*.

[90] P. Hacker, "The European AI liability directives—Critique of a half-hearted approach and *Comput. Law Secur. Rev.*, vol. 51, Nov. 2023, Art. no. 105871. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S026736492300081X

[91] R. Gozalo-Brizuela and E. C. Garrido-Merchan, "ChatGPT is not all you need. A state of the art review of large generative AI models," 2023, *arXiv:2301.04655*.

[92] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Adv. Neural Inf. Process. Syst.*, vol. 30. New York, NY, USA: Curran Associates, 2017, pp. 1–11. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[93] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[94] D. Lenat and G. Marcus, "Getting from generative AI to trustworthy AI: What LLMs might learn from cyc," 2023, *arXiv:2308.04445*.

[95] A. Baronchelli, "Shaping new norms for artificial intelligence: A complex systems perspective," 2023, *arXiv:2307.08564*.

[96] I. D. Raji, I. E. Kumar, A. Horowitz, and A. Selbst, "The fallacy of AI functionality," in *Proc. ACM Conf. Fairness, Accountability, Transparency*, Jun. 2022, pp. 959–972, doi: 10.1145/3531146.3533158.

[97] R. Gozalo-Brizuela and E. C. Garrido-Merchán, "A survey of generative AI applications," 2023, *arXiv:2306.02781*.

[98] M. Graziani, L. Dutkiewicz, D. Calvaresi, J. P. Amorim, K. Yordanova, M. Vered, R. Nair, P. H. Abreu, T. Blanke, V. Pulignano, J. O. Prior, L. Lauwaert, W. Reijers, A. Depeursinge, V. Andrearczyk, and H. Müller, "A global taxonomy of interpretable AI: Unifying the terminology for the technical and social sciences," *Artif. Intell. Rev.*, vol. 56, no. 4, pp. 3473–3504, Apr. 2023, doi: 10.1007/s10462-022-10256-8.

[99] W. Liang, G. A. Tadesse, D. Ho, L. Fei-Fei, M. Zaharia, C. Zhang, and J. Zou, "Advances, challenges and opportunities in creating data for trustworthy AI," *Nature Mach. Intell.*, vol. 4, no. 8, pp. 669–677, Aug. 2022. [Online]. Available: https://www.nature.com/articles/s42256-022-00516-1

[100] A. Balahur, A. Jenet, I. H. Torres, V. Charisi, A. Ganesh, C. B. Griesinger, P. Maurer, L. Mian, M. Salvi, S. Scalzo, J. S. Garrido, F. Taucer, and S. Tolan, "Data quality requirements for inclusive, non-biased and trustworthy AI: Putting science into standards," Joint Res. Centre (Eur. Commission), Publications Office Eur. Union, Brussels, Belgium, Tech. Rep. JRC131097, 2022, doi: 10.2760/365479.

[101] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 2014, *arXiv:1406.2661*.

[102] S. Nativi and S. De Nigris, "AI watch, AI standardisation landscape state play link to EC proposal for AI regulatory framework," Joint Res. Centre (Eur. Commission), Publications Office Eur. Union, Ispra, Italy, Tech. Rep. JRC125952, 2021, doi: 10.2760/376602.

[103] J. S. Garrido, S. Tolan, I. H. Torres, D. F. Llorca, V. Charisi, E. G. Gutierrez, H. Junklewitz, R. Hamon, D. F. Yela, and C. Panigutti, "AI watch: Artificial intelligence standardisation landscape update," Joint Res. Centre (Eur. Commission), Publications Office Eur. Union, Seville, Spain, Tech. Rep. JRC131155, 2023, doi: 10.2760/131984.

[104] (2021). *SAE J 3016—Taxonomy and Definitions for Terms Related to On-Road Motor Vehicle Automated Driving Systems*. SAE International, On-Road Automated Driving (ORAD) committee. [Online]. Available: https://www.sae.org/standards/content/j3016_202104/

[105] ASAM e.V. *ASAM OpenODD—Operational Design Domain*. Accessed: Sep. 10, 2023. [Online]. Available: https://www.asam.net/standards/detail/openodd/

[106] *AVSC Best Practice for Describing of an Operational Design Domain: Conceptual Framework and Lexicon*, Automated Vehicle Saf. Consortium, PA, USA, 2020.

GEORG STETTINGER received the B.Sc. and M.Sc. degrees in electrical engineering from the Graz University of Technology (TUG), Graz, Austria, in 2009 and 2011, respectively and the Ph.D. degree in information technology from the University of Klagenfurt (AAU), Klagenfurt, Austria, in 2015. From 2015 to 2018, he was a Senior Researcher with the Co-Simulation and Software Group, Virtual Vehicle Research GmbH, Graz, where he led the Control Systems Team, from 2018 to 2022. He is currently a Senior Project Manager with the Research and Development Funding Department, Infineon Technologies AG, Munich, Germany. His current research interests include ODD-based testing and validation, particularly the certification and homologation of automated vehicles. He is an active member of the technical committee of the driving simulator association with a special focus on virtual verification and validation of ADAS/AD systems.

PATRICK WEISSENSTEINER received the B.Sc. and M.Sc. degrees in mechanical engineering and business economics and the Ph.D. degree in electrical engineering from the Graz University of Technology, Graz, Austria, in 2015, 2017, and 2023, respectively. He is currently a Senior Researcher with the Automotive Electronics and Software Department, Virtual Vehicle Research GmbH, Graz. His current research interests include the safety validation of automated vehicles, with a particular focus towards coverage methods for the respective operational design domain of such vehicles.

SIDDARTHA KHASTGIR (Member, IEEE) is currently the Head of Verification and Validation for Connected and Autonomous Vehicles (CAV) and a Professor with The University of Warwick (WMG). He leads several collaborative research and development projects with industrial and academic partners nationally and internationally. His research interests include generating safety evidence and arguments, test scenario generation, simulation-based testing, and the safety of AI systems. Leveraging the cross-domain nature of safety, he is also involved in safety research in aviation, marine, and healthcare. He has been appointed as a member of the Department for Transport's Science Advisory Council. He is an active member of various national and international standardization and regulatory groups, including ISO, SAE, and ASAM. He also represents the U.K. on several ISO technical committees and he is the lead author for two new ISO standards for aspects of automated driving systems. He sits on the United Nations Economic Commission for Europe (UNECE) committees on safety of automated driving. Prior to joining WMG, he was with FEV GmbH, Germany, leading automotive software development and testing for series production projects. He has received numerous national and international awards for his research contributions, including the prestigious UKRI Future Leaders Fellowship, in 2019, focused on safety evaluation of CAVs.