embeddedworld2021 Exhibition&Conference ...it's a smarter world DIGITAL

Trustworthy AI-based Systems with VDE-AR-E 2842-61

Structured development for trustworthy autonomous/cognitive systems

Dr. Henrik J. Putzer
CEO at cogitron GmbH
Pliening near Munich, Germany
henrik.putzer@cogitron.de

Dr. Harald Rueß
Scientific and Managing Director
fortiss GmbH[1]
Munich, Germany
ruess@fortiss.org

Johannes Koch
VDE DKE
Frankfurt, Germany
johannes.koch@vde.com

***Abstract* — AI is used for implementing cognitive capabilities and increasingly autonomous computer systems. Despite technological advances, however, questions remain about the level of trust that can be placed in AI systems. Therefore, developers, assessors, and users of AI-based systems are currently in dire need for guidance and best practices for ensuring the trustworthiness of AI-based applications. These needs are addressed in the VDE-AR-E 2842-61 application rule for the *Development and trustworthiness of autonomous/cognitive systems*, which itself is based on the fundamental IEC 61508 industrial standard for safety engineering. In particular, the VDE-AR-E 2842-61 standard extends the established safety lifecycle of IEC 61508 to handle complex system-of-system approaches with complex behavior and AI-based components. We provide an overview of the structured approach of the VDE-AR-E 2842-61 and introduce its underlying methods and concepts for achieving and maintaining overall performance, intended behavior and trustworthiness of an autonomous/cognitive and AI-based system.**

Keywords — artificial intelligence, trustworthiness, safety, security, usability, system design, strcutured approach, process, standard, compliance, certification

I. MOTIVATIOIN

AI consists of a set of key technologies for fueling economic growth and for addressing a large number of important societal challenges. These technologies are often used for developing improved products and services, for analyzing and optimizing processes, and for the efficient control of machineries. The predictive capabilities of AI technologies may be also be used for predictive maintenance of machineries or for detecting and predicting malicious attacks. In addition, AI-based software assistants are getting increasingly popular for implementing automated and improved decision-support systems and for supporting, say, improved and AI-aided diagnosis in medicine. AI technologies are also used for implementing increasingly autonomous and safety-critical systems such as self-driving cars, swarms of service drones, surgical robots, and for controlling societal-scale service infrastructures such as energy or water distribution.

The technological advances driven by AI is grounded in new characteristics and features of that new technology. For example they can implement implicit requirements by learning from example. Furthermore, AI-based systems are can be designed to learn continuously, adapt, and optimize themselves based on experience, operate partially in unknown or uncertain environments. On the downside they are incompatible with complete traceability of requirements and consequently have problems with classical verification. AI systems and their generated behavior offer a variety of new attack surfaces (e.g. sensor spoofing), and often lead to largely unpredictable and emergent behavior in operation.

Despite technological advances that have led to a proliferation of AI-based solutions, questions remain about the level of trust that can be placed in such software systems, especially when considering the downsides above. What is missing, in particular, is a rigorous and structured approach to building and operating AI systems in which people can trust. In particular, attributes of trustworthiness including functional safety, cybersecurity, privacy, usability and maintainability as well as legal and ethical aspects are relevant.

In traditional safety engineering, the basis for certification is an assurance case. This is a structured and convincing argument for the trustworthiness of the system under consideration – with respect to a well-defined operating environment, for pre-defined use cases and for the intended purpose and benefit. Such a

[1] The research was supported by the *Robust AI* project at fortiss as funded by the Bavarian Ministry of Economics

structured argument with its evidences is closely correlated to a structured development approach delivering (in a structured, documented and reproducible manner) the system under consideration together with suitable development artefacts (e.g. design reviews, test reports).

For technologies such as electronics and software structured development approaches (processes, methods, artefacts) are prescribed through industrial standards such as the domain-agnostic IEC 61508 and its domain-specific instantiations such as ISO 26262.

Currently there is no such structured approach for developing technical systems based on AI. There is no generally accepted and documented development approach nor is there a generally accepted and documented way to ensure trustworthiness when it comes to the development of AI-based systems. Moreover, there is not yet a relatively complete and continuous set of generally accepted methods and tools for supporting the complete life-cycle for engineering AI-based systems.

The VDE-AR-E 2842-61 *Development and trustworthiness of autonomous/cognitive Systems* takes up these challenges for engineering trustworthy AI (see [1], [2], [3], [4], [5], and [6]). This standard[2] describes a generic framework for the development of trustworthy solutions and trustworthy autonomous/cognitive systems. It defines a reference lifecycle in analogy to the key functional safety standards (i.e. [7]) as a unified approach to achieve and maintain the overall performance of the solution and the intended behavior and trustworthiness of the autonomous/cognitive system.

The VDE-AR-E 2842-61 has been developed by the working group DKE AK 801.0.8 of the VDE Association for Electrical, Electronic and Information Technologies under the direction and sustainable contribution of fortiss, cogitron and other relevant industrial and research partners.

Here, we present the basic approach and key concepts underlying the development of the VDE-AR-E 2842-61 standard. We also describe the current status and upcoming developments.

II. Related Work

A number of research groups address the verification of AI-based systems. Architectural approaches (see [8]), for example, measure the related uncertainty or potential functional disfunction or approach to directly influence the behavior of AI. However, these approaches usually are incomplete in that they not address the systems engineering perspective, which includes the operating environment, users and other stakeholders via a systems-engineering process down to the AI-elements and its development (see *related work* in [9]).

A more holistic approach is usually being taken when developing industrial engineering standards. There are currently a number of committees for proposing and developing structured AI developments. A rather comprehensive list of such on-going standardization efforts, in particular, the working groups at the International Standardization Organization (ISO) within the subcommittee ISO/IEC JTC1 SC42, is described in [10].

The VDE in the DKE AK 801.0.8 has been developing a standardized framework for trustworthy AI since October 2017. This application rule captures the state-of-the-art and science in that is based on current research (e.g. [11]) and the comprehensive experience of its members in developing AI-based systems. The result of this effort is the VDE-AR-E 2842-61 *Development and trustworthiness of autonomous/cognitive Systems*.

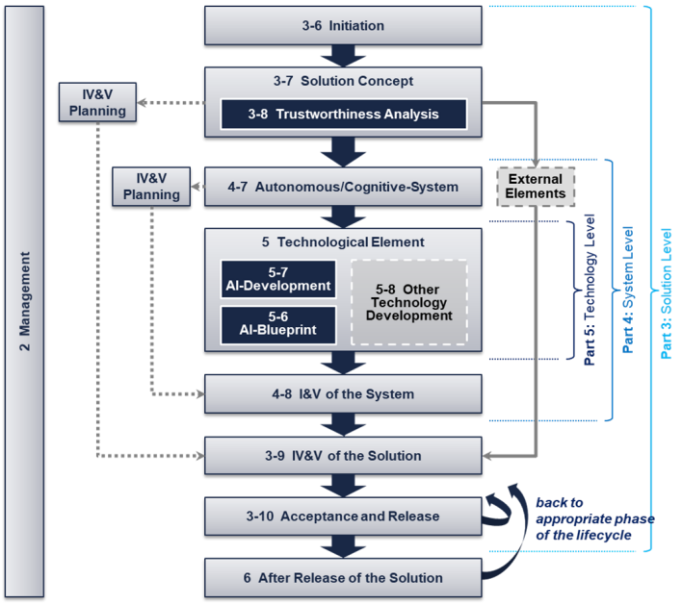

Figure 1: Trustworthiness Reference Lifecycle

III. The Approach

Our approach for defining the VDE-AR-E 2842-61 is based on three initial thoughts:

1. *AI = new technology & new engineering approach*
   Even if AI in some cases appears to be magic, in science and engineering we do not deal with magic. AI is "just" a new engineering approach, a new technology with new characteristics (that are even by AI scientists not fully understood) and its specific development approach (e.g. engineering method) that produces good old automation with all the known problems as described in [21].
2. *AI is (only) one element within a system or product*
   AI never stands alone. It is always a component within the context of a system which in turn is situated within the context of an environment that needs to be considered. Hence, an adopted *systems-engineering* process needs to be considered for the development of AI based systems.

[2] The VDE-AR-E 2842-61 *Development and trustworthiness of autonomous/cognitive Systems* actually is an application rule (*ge.: Anwendungsregel*) which can, due to public commenting procedure, become acknowledged rule of technology ('state-of-the-art' in Germany.

3. *Need for a structured approach*
   Key to trustworthy AI is to provide an appropriate *assurance case* based on a structured argumentation and referencing evidences. Such a structured argumentation needs to rely on a structured development approach (e.g. process). Artifacts like test or verification reports serve as evidences in that argumentation.

These initial thoughts lead to the outline of our approach: based on modern systems-engineering we need a structured approach that considers trustworthiness aspects and new technologies to develop trustworthy autonomous/cognitive systems that are based on AI. To accomplish this, we us the ICE 61508 (see [7]) as a starting point. This international standard is the industry independent and generic master of all standards handling functional safety in electric, electronic and programmable systems (E/E/EPS systems).

Basically the IEC 61508 defines a risk-based approach along a reference lifecycle with a structured approach (process) including requirements on measures and methods. This is the approach adopted and extended by the VDE-AR-E 2842-61 (see Figure 1). The most relevant extensions and new key concepts are discussed in the next section.

IV. New Key Concepts and Extensions

The backbone of the VDE-AR-E 2842-61 is the risk-based approach along the trustworthiness reference lifecycle (see Figure 1), which is derived from the lifecycle in [7]. The main extensions and used key concepts can be described as follows:

A. Autonomous/cognitive System

The term autonomous/cognitive system (A/C-system) is a new term, a made-up word coined by the VDE-AR-E 2842-61. It denotes the special characteristic of complex systems in complex environments, trustworthiness aspects and potentially but not necessarily the use of AI. Furthermore it takes into account the common use of *autonomous* in contrast to the scientifically correct term *fully-automated*.

We are using the term A/C-system as a term for (technical) systems that bear behavior one would usually associate with human behavior in terms of complexity and in terms of abstract description, e.g. using mental terms like belief, goal or planning. Such a system may or may not make use of artificial intelligence (AI). The term is associated with the special kind of behavior. This behavior takes the situation into account and is based on decision the system takes.

The term A/C-system conceptually matches the term E/E/PE system in the IEC 61508 or the item under consideration in ISO 26262 but it is to be understood in a much broader scope. The A/C-system is the system that shall be developed applying the VDE-AR-E 2842-61 (e.g. a fully-automated car, a medical robot or an internet service).

B. From safety to trustworthiness

The IEC 61508 deals with *functional safety*. Currently a lot of aspects have to be handled during modern system development: safety (incl. functional safety and safety of the intended function), security, privacy, usability, ethics, etc. In the VDE-AR-E 2842-61 this leads to the meta term *trustworthiness* (see [1]) which actually is the per-project suitable selection and combination of aspects as indicated in Figure 2.

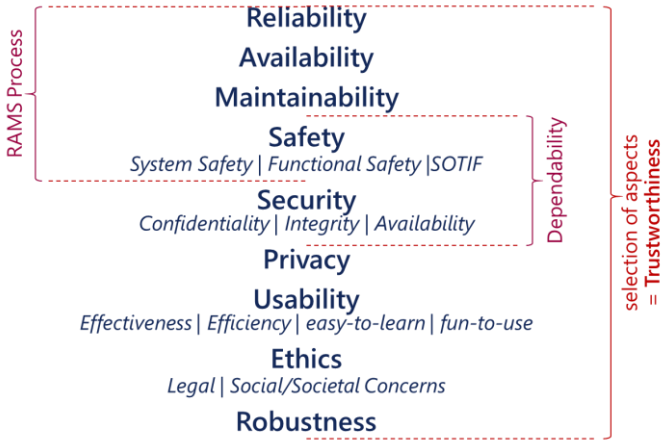

Figure 2: Aspects of Trustworthiness

C. Solution Level

The *solution level* (see [3]) adds one level of abstraction above the A/C-system which results in an additional phase in the trustworthiness reference lifecycle. The solution includes the AI-system as a black box and examines its role and behavior in the overall environment including the user, other interface partners and stakeholders. The solution level contains the ideas on the socio-technical work system including all modelling and analyses.

The VDE-AR-E 2842-61 proposes an approach of modelling the solution level as a system, e.g. using SysML. In this model at the solution level, the so-called *solution concept*, the A/C-system is one element which should be described as a black box (*black box model* of A/C-system) and with a *white box model* which is rather functional description how the A/C-system generates its behavior based on a certain cognitive theory or the *sense-plan-act* architecture (see [2], section 12).

This solution level is the origin of all hazards. Approaches like STAMP and STPA (see [15]) are compatible with this level. To identify and quantify all hazards well known analyses of AI trustworthiness aspects are executed (e.g. hazard analysis and risk assessment – HARA or thread analysis and risk assessment - TARA) resulting in a list of hazards with heterogeneous attributes (e.g. fault tolerant time interval, safe state, SIL). These hazards are mitigated by *trustworthiness measures* that are attributed with harmonized hazard attributes and with the *trustworthiness performance level* (TPL). These trustworthiness measures are allocated to elements in the refined solution concept, which now is addressed as trustworthiness solution concept as major result of the solution level.

D. System Level

The *system level* (see [4]) represents a modern systems-engineering approach. All defined activities and requirements are iteratively applied to form the hierarchical development of the A/C-system. Each iteration – from A/C-system via sub-system, component, sub-component etc. – refines the requirements and architecture of the design and carefully keeps

track of the trustworthiness measures and detailed trustworthiness functions and requirements (incl. TPL).

The system level provides means for TPL inheritance including allocation (*vertical inheritance*), decomposition (*shared inheritance*), and independence (or *horizontal inheritance*). Additionally, it provides design *patterns* to support the verification of AI-based systems and the implementation of certain AI features.

The overall purpose of the system level is to link the abstract level of the solution concept to the detailed level of the technology (e.g. hardware/electronics, software and artificial intelligence).

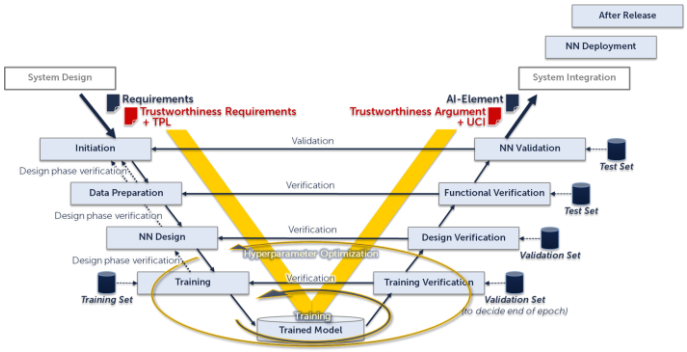

Figure 3: Example AI-blueprint for neuronal networks

E. Technology Level and the AI-blueprint

The purpose of the *technology level* (see [5]) is to define the actual implementation of the elements that have been defined by the system level during its last iterative application. This implementation description needs to consider "all kinds of technology" including but not limited to AI, hardware, software, etc. This is handled in two basic steps:

In a *first step*, AI technologies are separated from all other technologies. These *other technologies* are handled through an interface to already existing and well trusted standards that are able to handle these classical other technologies (e.g. for the development of classical software there are standards like IEC 61508-3, ISO 26262-6 or IEC 62304 that already provide approved approaches).

The *second step* is about handling all AI technologies. To cover the special characteristics of *any* AI technology the term *AI-blueprint* is introduced: An AI-blueprint can be regarded as a small structured process definition to develop an AI element using the respective AI technology (including all necessary measures, methods and good practices). Three sections of [5] are dealing with AI-blueprints:

- section 5-6 provides requirements on how to define and qualify am AI-blueprint,
- section 5-7 provides requirements on how to apply such an AI-blueprint, and
- section 5-9 (as a non-normative appendix) provides ready to use examples of AI-blueprints.

Thus, an AI-blueprint is a generic and structured approach for the development of an AI element based on a certain AI technology. It includes a clear development contract including trustworthiness related assumptions and guarantees to ensure a seamless "plug-in" integration into the overall trustworthiness related development process of the VDE-AR-E 2842-61. An example for such an AI-blueprint scoping the development of deep neuronal networks is given in Figure 3.

F. Uncertainty Confidence Indicator (UCI)

Each type of technology has its own types and causes of failures. Current standards like the IEC 61508 propose that software has *systematic failures*, only. Measures to avoid such systematic failures include a good development culture (e.g. safety culture), relying on experts and well-known designs, methods and measures ideally defined in a documented process. For electronic elements we see *random failures* as an additional further type. Quantitative measures (e.g. based on fault rates and fault tree analyses) and metrics like *safe-failure-fraction* or the *diagnostic coverage* help to develop safe (or even trustworthy) designs.

With some AI technologies (e.g. neuronal networks) we see a third type of failure: the *uncertainty-related failure*. This failure cannot be mitigated by good processes and established metrics. It is a characteristics and new kind of failure that is inherent to the technology of neuronal networks and some other machine learning approaches (see [5]). To handle this third kind of failure the *uncertainty confidence indicator* (UCI) is introduced (see Figure 4 and [5]).

type of failure	measures	HW measures	SW measures	AI measures
systematic	qualitative requirements	systematic capability	systematic capability	systematic capability
random	quantitative requirements	λ, SFF, DC, target values	-- / --	-- / --
uncertainty-related	structured approach	-- / --	-- / --	Uncertainty confidence indicator (UCI)

Figure 4: Three Types of Failures plus Mitigating Measures

G. Trustworthiness Assurance Case

Finally, the structured approach in the VDE-AR-E 2842-61 proposes the *trustworthiness assurance case* (see [3], section 13). Based on scientific research (see [17] and [18]) this assurance case considers all trustworthiness aspects, the risk-based approach and the overall sound argumentation that the AI system is trustworthy in the defined use cases and environments. This argumentation is based on evidences that are derived from development artifacts generated during the development process (e.g. design verification reports, test reports).

The trustworthiness assurance case with its structured argumentation is supposed to be used in order to achieve at least two goals:

1. Traditionally the trustworthiness assurance case helps to structure the trustworthiness argument. It is the easy entry point to understand a project, suitable for contributors or assessors/auditors.
2. As a planning tool the argumentation in the trustworthiness assurance case is a supplement of the

project plan: the project plan cares about time and milestones while the argumentation is a plan regarding content and semantic correlation of activities and artifacts. It can be used to determine the essential documents and analyses that are necessary for the final argumentation. When it comes to tailoring the efforts, the priorities can be derived from the argumentation.

V. Current State and Results

The VDE in the working group DKE AK 801.0.8 is working on the VDE-AR-E 2842-61 *Development and trustworthiness of autonomous/cognitive Systems* since October 2017. It is based on several research papers (e.g. [11]) and gained maturity over time (see [12], [13], and [14]).

The VDE-AR-E 2842-61 now is the first available standard that combines expert knowledge, practical experience from industry, the latest research in AI and profound standardization expertise to cover the development of AI-based systems. It clearly separates ethical fundamentals and societal acceptability from the technical approach (see Figure 5 for all major characteristics).

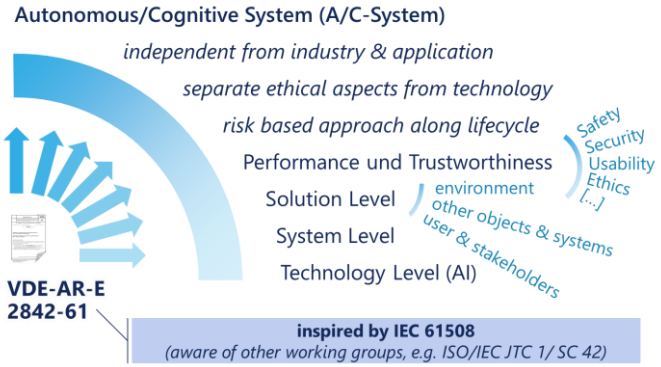

Figure 5: Characteristics of VDE-AR-E 2842-61

The VDE-AR-E 2842-61 provides a generic framework for the development of trustworthy solutions and trustworthy autonomous/cognitive systems. It defines a reference lifecycle in analogy to the key functional safety standards (i.e. IEC 61508) as a unified approach to achieve and maintain the overall performance of the solution and the intended behavior and trustworthiness of the autonomous/cognitive system.

The VDE-AR-E 2842-61 consists of six parts (see [1], [2], [3], [4], [5], [6]) plus additional parts containing application guides (see Figure 6). As of end of January 2021 parts 1, 2, 3 and 6 are publicly available. Parts 4 and 5 are in their final development phase and are planned to be available soon. The application guides have been postponed.

So far there is no overall industrial and practical application of that standard. First scientific applications as in [19] delivered both: On one side we got promising results in structuring the development of autonomous/cognitive systems, providing a framework to contextualize modern systems-engineering approaches and AI related methods and measures. On the other side many questions arose concerning details in process interfaces, methods and application practice (see next section *VI Conclusion and Future Work*).

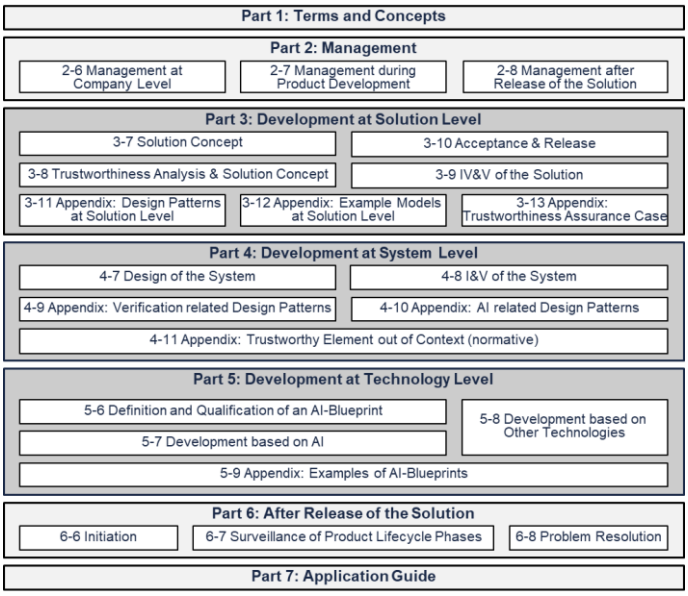

Figure 6: Graphical overview of VDE-AR-E 2842-61

Meanwhile reading the VDE-AR-E 2842-61 and working with it can generate benefits in many scenarios:

- The overall application in new projects dealing with AI technologies correlated with safety or any other trustworthiness aspect will improve a structured, well planned and traceable approach.
- Even an application of the standard in parts makes sense: the innovative risk-based approach together with a modern systems-engineering and the requirements on assurance case will be a great structuring momentum on complex systems design.
- If the standard is not applied directly it still can be used as a source for checklists targeting trustworthiness assessments for already existing design or even for certification and/or homologation efforts.

VI. Conclusion and Future Work

The VDE-AR-E 2842-61 describes and determines the state-of-the-art in structured development of trustworthy AI-based systems. Consequently, this application rules provides the framework for developing AI-based products with a clear perspective of getting them certified for the market. The VDE-AR-E 2842-61 answers the question of how to build AI, how to verify AI (incl. trustworthiness assurance case) for developers, and provides a reference framework for how to certify AI-based systems.

Although the VDE-AR-E 2842-61 is available as a stable and rigid framework for the development of A/C-systems, there are some topics for future evolutions of the standard:

apply ~ eval ~ improve: Now that the VDE-AR-E 2842-61 is available it should find wide-spread application in a large number of industrial projects from various application domains. Further experience with applying this development framework yields new insights which is expected to be included in further editions of the VDE-AR-E 2842-61 application rules. In

particular, further experience in developing these kinds of systems will lead to more detailed tool boxes for metrics, development methods, and verification measures.

However, in updating this application rule we still need to ensure that the standard remains applicable with small overhead so that e small and medium companies (SMEs) are also able to apply it in their product developments.

Research on AI: The content of the VDE-AR-E 2842-61 needs to be refined. A lot more detail is necessary to improve the measures, methods and metrics on AI technologies. And more breadth is necessary, e.g. more AI-blueprints to cover more AI technologies beyond the content of [5]. Further research is needed to better understand AI technologies like neuronal networks and to understand how trustworthiness can be measured (UCI), ensured and included in an assurance case. In addition, further topics such as continuous integration, high number of variants in products and re-usability need to be supported.

Integrate knowledge from other working groups: VDE/DKE is the first one to come up with a standard on structured development of trustworthy AI-based systems, but certainly they are not the only ones. Many groups are working on this topic including ISO/IEC JTC1 SC42 (see extended list in [10]). It seems to be worthwhile of synchronizing the ideas and solutions of these initiatives into one, hopefully, harmonized, structured approach for trustworthy AI.

Internationalization: The VDE-AR-E 2842-61 is a national standard. The goal, however, is to internationalize the standard through ISO or IEC. Besides the activities to integrate and harmonize the knowledge from other working groups, there are currently activities by, among others, Japan of adopting the VDE-AR-E 2842-61 also as their respective national standard on trustworthy AI.

Certification: When it comes to certification and homologation of products the VDE-AR-E 2842-61 serves as a reference model. Questions can be answered like *What is the certification interface between developers and certification?* and *How to proof test AI based systems with safety-relevance or trustworthiness requirements?* - In the long run specific AI related certificates can be developed. Such certificates help less experienced people and users to make buying decisions and to establish trust on AI based products. Overall these certificates should enhance acceptance of and confidence in AI-based products, thereby leveraging economic success of AI-based products and services.

VII. Call for Participation

Finally, if you and/or your organization are interested in discussing any topic related to the VDE-AR-E 2842-61, in contributing to the evolution of it, in applying this standard to your development project, or in pushing the state-of-the-art in autonomous/cognitive systems-engineering further, please contact us– we are more than happy in supporting you and your case of AI.

References

[1] VDE-AR-E 2841-61-1:2020, “Specification and Design of autonomous / cognitive Systems - Part 1: Terms and Concepts”;

[2] VDE-AR-E 2841-61-2:2020, “Specification and Design of autonomous / cognitive Systems - Part 2: Management”;

[3] VDE-AR-E 2841-61-3:2020, “Specification and Design of autonomous / cognitive Systems - Part 3: Development at Solution Level”;

[4] VDE-AR-E 2841-61-4:2021 (unpublished), “Specification and Design of autonomous / cognitive Systems - Part 4: Development at System Level”;

[5] VDE-AR-E 2841-61-5:2021 (unpublished), “Specification and Design of autonomous / cognitive Systems - Part 5: Development at Technology Level”;

[6] VDE-AR-E 2841-61-6:2020, “Specification and Design of autonomous / cognitive Systems - Part 6: After release of the Solution”;

[7] IEC 61508:2010-04 “Functional safety of electrical/ electronic/ programmable electronic safety-related systems”, 2nd Edition, April 2010

[8] Chih-Hong Cheng, Dhiraj Gulati and Rongjie Yan: “Architecting Dependable Learning-enabled Autonomous Systems: A Survey”. arXive, ID 1902.10590, 2019

[9] Putzer, H.J. and Wozniak, E., 2020. A Structured Approach to Trustworthy Autonomous/Cognitive Systems. arXiv preprint arXiv:2002.08210.

[10] Wahlster W., Winterhalter C. (Herausgeber): “Deutsche Normungsroadmap Künstliche Intelligenz“, DIN & DKE, 2020-11

[11] H.Putzer, E.Wozniak: „Trustworthy Autonomous/Cognitive Systems – A Structured Approach”, white paper, https://www.fortiss.org/veroeffentlichungen/whitepaper, 2020-10

[12] H.Putzer: “Ein strukturierter Ansatz für verlässliche KI”, Key-Note, VDE-DKE-Tagung Funktionale Sicheheit für die Zulunft, 2019-03, Erfurt

[13] H.Putzer: “Ein Referenzmodell für vertrauenswürdige KI: Vorstellung eines neuen VDE-Standards - Vorstellung“, VDE tec summit, Berlin 2020-02

[14] H.Putzer: “Ein Referenzmodell für vertrauenswürdige KI: Vorstellung eines neuen VDE-Standards – Embedded Systems“, VDE tec summit, Berlin, 2020-02

[15] N. Leveson: “Engineering a Safer World: Systems Thinking Applied to Safety”. Cambridge: The MIT Press, 2011

[16] Döring, B.: “Systemergonomie bei komplexen Arbeitssystemen”, in: Hackstein, R. et. al. (Hrsg.): Arbeitsorganisation und Neue Technologien. Impulse für eine weitere Integration der traditionellen arbeistwissenschaftlichen Entwicklungsergebnisse. Berlin, Heidelberg, New York 1986, page 399-434

[17] Wozniak, E., Cârlan, C., Acar-Celik, E. and Putzer, H.J., 2020, September. A Safety Case Pattern for Systems with Machine Learning Components. In International Conference on Computer Safety, Reliability, and Security (pp. 370-382). Springer, Cham.

[18] Wozniak, E., Putzer and Cârlan, C., 2021. AI-Blueprint for Deep Neural Networks. In SafeAI@ AAAI.

[19] H.Putzer, V.Nigam, Th.Brunner: “Dependable Autonomous/Cognitive Systems”, VDA Automotive Sys, Potsdam, 2019-06

[20] H.Putzer, H.Rueß, J.Koch: „ Trustworthy AI-based Systems With VDE-AR-E 2842-61 “, (in press, ID10334) embedded world 2021, online, 2021-03

[21] Billings, C.E: “Human-Centered Aviation Automation: Principles and Guidelines”, Technical Memorandum, NASA TM, 1996