Weshalb die VDE-AR-E 2842-61 (vertrauenswürdige KI-Systeme) nicht nur die Entwicklung betrifft

13. April 2021 

Expert:
Prof. Dr. Christian Johner

Der VDE hat mit der **VDE-AR-E 2842-61** eine ganze Familie an normativen Vorgaben für vertrauenswürdige autonom kognitive Systeme wie z.B. **KI-Systeme** erarbeitet. Obwohl diese „Anwendungsregeln" nicht spezifisch für eine Domäne wie z.B. Medizinprodukte sind, stellen sie dennoch eine Fundgrube für viele Medizinprodukthersteller dar.

Dieser Artikel zeigt Ihnen,

- was KI-Systeme sind,
- welche Hersteller welche Teile dieser Normenfamilie berücksichtigen sollten,
- welche konkreten Anregungen diese Teile enthalten und
- weshalb die VDE-AR-E 2842-61 keinesfalls nur die Entwicklung die betrifft.

1. Vertrauenswürdige KI-Systeme

a) Definition und Abgrenzungen

KI-Systeme sind eine Untermenge der autonom kognitiven Systeme, die Verfahren der künstlichen Intelligenz verwenden.

Autonome kognitive Systeme sind wie folgt definiert:

> Ein System wird als autonom (kognitiv) bezeichnet, wenn es ohne menschliche Steuerung oder detaillierte Programmierung ein vorgegebenes Ziel

autonom/kognitives System, A/K System)

„is a technical system that is able to generate autonomous and cognitive behavior. Within the context of this AR an A/C-system is part of a solution.

The term autonomous/cognitive system and especially the germen variant "autonom/kognitives System" is a new term, a made-up word coined by this AR. It denotes the special characteristic of complex systems in complex environments (covered by the solution level), trustworthiness aspects and potentially but not necessarily the use of AI in one or more elements of the system. Furthermore it takes into account the common use of "autonomous" in the public along with the expectations on complex behavior of such systems (e.g. in a shop one would rather order an "autonomous car" than a "fully automated car")."

VDE-AR-E 2842-61-1, Abschnitt 3.1.8

Die Norm definiert die Eigenschaft „**autonom**" nicht ganz genauso. Aber auch im Sinne der Norm bedeutet „autonom" „ohne menschliche Steuerung". Die Norm arbeitet zusätzlich mit dem Begriff „kognitiv" und „kognitive Schleife" um das situationsspezifische Verhalten zu beschreiben.

Weil es sehr viele Situationen gibt, auf die solche Systeme reagieren können müssen, ist meist eine „detaillierte Programmierung" nicht möglich. Daher verwenden viele autonom kognitive Systeme Verfahren des maschinellen Lernens als Teilgebiet der künstlichen Intelligenz.

Umgekehrt sind aber nicht alle Systeme, die KI verwenden, auch KI-Systeme und damit autonom kognitive Systeme. Beispielsweise ist eine Software, die mit Hilfe der KI Krebs auf einem CT-Bild erkennt, kein KI-System im Sinne dieser Definition.

Abb. 1: Abgrenzung von autonom kognitiven Systemen, KI-Systemen und Produkten, die KI nutzen.

Weiterführende Informationen

Erfahren sie in dem Artikel zu den autonomen Systemen, was deren spezifischen Vorteile und Risiken sind und welche regulatorischen Anforderungen zu beachten sind.

Dieser Artikel nutzt im Folgenden den Begriff „KI-System“, da dieser der populärere Begriff ist und die „KI-Systeme“ in den Anwendungsbereich der VDE-AR-E 2842-61 fallen.

b) Beispiele für KI-Systeme

Beispiele für KI-Systeme sind:

- Desinfektionsroboter wie der von XENEX
- Roboter, die Aufgaben in medizinischen Laboren erledigen wie der von ABB
- Pflegeroboter und weitere Roboter in Krankenhäusern wie z.B. autonome kooperative OP-Roboter, die selbständig OPs ganz oder teilweise durchführen und ggf. zukünftig sogar nur durch Laien bedient werden (z.B. in entlegenen Gebieten)
- Künstliche digitale Pankrease

c) Vertrauenswürdigkeit

Die Vertrauenswürdigkeit ist hier als Metabegriff zu verstehen, der die Safety, Cybersecurity, Effectiveness, Usability etc. umfasst.

> Trustworthiness [...] combines several aspects of trustworthiness in a quite generic way: for every product the set of aspects can be suitably selected and remains unchanged throughout the project. Aspects of trustworthiness include but are not limited to system safety, functional safety, safety of use, security, usability, ethical and legal compliance, reliability, availability, maintainability, and (intended) functionality.

Abb. 2: Aspekte der Vertrauenswürdigkeit („Trustworthiness“)

2. Spezifische Risiken von KI-Systemen

a) Risiken durch autonom kognitive Systeme (im Allgemeinen)

Der Artikel zu den autonomen Systemen hat bereits einige Risiken genannt, die für diese Klasse an Systemen spezifisch sind. Dazu zählen Risiken durch

- die Autonomie der Systeme,
- verschieden technische Kontexte,
- verschiedene klinische Kontexte,
- adaptive Algorithmen und
- mangelnde Interoperabilität.

Bei KI-Systemen kommen spezifische Risiken hinzu, die die folgenden Kapitel vorstellen.

b) Risiken durch unzureichende Zweckbestimmung

Hersteller müssen in der Zweckbestimmung konkret festlegen, für welche

- Nutzer,
- Nutzungsumgebungen,
- Patienten,
- Krankheiten, Diagnosen, Kontraindikationen,

zu bestimmen.

c) Risiken durch unbekannte Situationen

Gerade weil es sehr schwer ist, alle Situationen vorherzusagen, gelingt es den Herstellern nicht immer, vollständige Spezifikationen zu erstellen.

Selbst wenn die Hersteller eine Situation antizipieren, ist es häufig herausfordernd, für jede das optimale Systemverhalten zu spezifizieren.

Ohne diese präzisen Spezifikationen und Produktanforderungen wird es den Entwicklungsabteilungen und Data Scientists nur schwer gelingen, spezifische Anforderungen an die KI-Modelle und an das Sammeln von Daten für deren Training abzuleiten.

Falls bei der Spezifikation und Entwicklung eines Produkts eine Situation nicht vorhergesehen wurde, ist auch das Verhalten des Produkts in dieser Situation nicht immer vorhersehbar.

d) Risiken durch Spezifikationslücken

Anforderungen müssen auf allen Abstraktionsebenen der Entwicklung klar und spezifisch dokumentiert werden. Dies gilt auch für KI-Systeme und die enthaltenen KI-Komponenten.

Insbesondere zur Entwicklungszeit selbstlernende KI-Komponenten verleiten dazu, die Spezifikation unklar zu gestalten, da ja diese KI-Komponente „schonen lernen wird, was sie zu tun hat“. Aber genau das ist ein Missverständnis: KI-Komponenten können lernen, „wie“ sie etwas machen, nicht aber das „was“.

KI-Komponenten dürfen nicht als Sammelbecken für unklare oder ungenaue Spezifikation dienen. Damit wäre die Entwicklung zum Scheitern verurteilt. Daher ist eine klare Traceability der Anforderungen einschließlich der Trustworthiness-Attribute (Performance, Safety, Security, Usability, ...) vorzusehen.

e) Risiken durch ungeeignetes Training der Modelle

Alle handwerklichen Fehler beim Entwickeln von KI-Modellen können auch bei KI-Systemen zu Risiken führen wie beispielsweise:

f) Risiko durch mangelnde Gebrauchstauglichkeit

Gerade bei KI-Systemen kann eine unzureichende Gebrauchstauglichkeit zu besonders hohen Risiken führen. Dieser Aspekt der „Human Factors“ führt sogar zu einer „Irony of Automation”. Dazu zählen:

- Ungerechtfertigtes Vertrauen in die Automation
- Ungerechtfertigtes Misstrauen in die Automation
- Unkenntnis der Grenzen der Automation
- Unfähigkeit zur Entscheidung, wann der Mensch eingreifen muss
- Höhere statt niedrigerer Komplexität

g) Risiken durch böswillige Nutzung

Deep fakes sind nur ein Beispiel, wie die KI missbraucht werden kann. Auch bei Medizinprodukten, die Verfahren der KI anwenden, wurde zumindest im Labor gezeigt, wie Systeme zur Klassifizierung von Bildern in die Irre geführt oder zur Preisgabe sensibler (Trainings-) Daten veranlasst werden können.

3. VDE-AR-E 2842-61: Vertrauenswürdigkeit von KI-Systemen gewährleisten

Die Anwendungsregel VDE-AR-E 2842-61 möchte dazu beitragen eben diese Risiken zu beherrschen und auf diese Weise die Vertrauenswürdigkeit von KI-Systemen zu gewährleisten.

Dabei hat die VDE-AR-E 2842-61 den Anspruch, den ganzen Produktlebenszyklus von der Produktidee bis zur der Phase abzudecken, die in der Medizinproduktewelt als „Post-Market Surveillance“ bezeichnet wird.

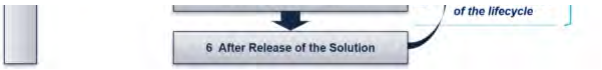

Abb. 3: Die Anwendungsregel VDE-AR-E 2842-61 besteht aus mehreren Teilen, die den ganzen Lebenszyklus von KI-Systemen abdecken. Die Nummern beziehen sich auf die Teile / Bände dieser Anwendungsregel. (Quelle: Dr. Henrik Putzer, Arbeitskreis *VDE-AR-E 2842-61*) (zum Vergrößern klicken)

a) „Initiation“: Festlegen der Zweckbestimmung

Die VDE-AR-E 2842-61 bietet mehrere Ansätze an, um bei KI-Systemen mit den Risiken durch unvollständige oder unklare Zweckbestimmungen umzugehen.

Problem	**Lösungsansätze der VDE-AR-E 2842-61**
Für generische Systeme wie Cobots (siehe oben) gibt es keine spezifische Zweckbestimmung. Diese können einfach für neue Zweckbestimmungen umprogrammiert oder angepasst werden.	Konzepte für **generische Sicherheitsnachweise** („Trustworthiness out of context“) in Anlehnung an die Automotive Norm ISO 26262
Keine klare Erwartungshaltung, sondern Ansicht „die KI löst dann alles schon irgendwie intelligent“	Festlegung von Use Cases und dem zu erzielenden Nutzen (intended benefit). Diese werden mit einer klaren **Ontologie** hinterlegt, die auch bei der Anforderungsbeschreibung (inkl. Traceability) genutzt und in späteren Phasen verfeinert und nutzbar gemacht wird bis hin zu Coverage-Metriken der Datensätze.
Das KI-System wird nicht über seinen ganzen Lebenszyklus betrachtet. Beispielsweise fehlen Elemente des bestimmungsgemäßen Gebrauchs wie das Update und die Wartung des Systems.	Produkt-Lebenszyklus mit Hilfe einer **Customer Journey Map** bzw. **UX / Experience-Map** modellieren

Unpräzise und nicht eindeutige Spezifikationen	▪ BPMN/SysML für Beschreibung der Black-box ▪ Abnahmekriterien, die die Performance und alle „Trustworthiness-Aspekte“ abdecken. Diese Abnahmekriterien geben auch die zu erreichenden Abdeckungsgrade an, die bei den Tests erzielt werden müssen
Die Aufteilung der Aufgaben und Verantwortlichkeiten zwischen Anwendern und KI-System ist nicht präzise festgelegt (z.B. zwischen dem Chirurgen und OP-Roboter)	**Definierte Notation** (z.B. BPMN/SysML), um ein „Solution Concept“ für die Blackbox des Systems zu modellieren.
Spezifische Risiken durch unterschiedliche Situationen z.B. Verfügbarkeit von Komponenten des Systems	▪ Dynamisches Risikomanagement wie im Artikel zu den autonomen Systemen beschrieben. ▪ mehrdimensionale Risikoanalyse (trustworthiness = Safety + Security + Usability + Ethik + … hier: safe + effective + secure) und Definition von konfliktfreien Entwicklungszielen (= trustworthiness goals) zur Abdeckung aller Risiken
KI wird als Platzhalter für unklare Funktionalität oder technische Umsetzung genutzt	Funktionsmodell des KI-Systems auf Basis von sense-plan-act oder einer anderen Kognitiven Theorie formulieren. Das liefert in der nächsten Phase auch das „white-box“-Modell des .

c) „System Level“: Beim Entwerfen der System-Architekturen von KI-Systemen

Auch für die Entwicklung von KI-Systemen bietet die VDE-AR-E 2842-61 Lösungsansätze für typische Risiken.

Problem	Lösungsansätze der VDE-AR-E 2842-61

Gewählte Architektur ist nicht die „beste“	Design Patterns anwenden für den Nachweis der Trustworthiness und zur Erreichung bestimmter KI-relevanter Produkteigenschaften (kontinuierliches lernen, Erklärbarkeit, etc.)

d) Beim Machine Learning (von Daten spezifizieren bis Modelle trainieren)

Problem	Lösungsansätze der VDE-AR-E 2842-61
Sicherheitsnachweise sind schwerer zu führen	▪ Fortschreiben der „Trustworthiness Assurance Case“ wie z.B. hier für KI beschrieben. ▪ Verwenden von Metriken für diese Assurance Cases wie z.B. Dichte von Datensätzen, Heat-Map-Analysen, Adversarial Sensitivity
Suboptimales Modell gewählt	▪ Verwenden von KI-Blueprints ▪ Beschränken des Einsatzes der KI

e) Beim Verifizieren und Validieren der KI-Systeme

Die VDE-AR-E 2842-61 nennt Lösungsansätze für alle Ebenen der Verifizierung und Validierung.

Problem	Lösungsansätze der VDE-AR-E 2842-61
Falsche Schlussfolgerungen aus Testergebnissen z.B. weil im Trustworthiness Assurance Case ein Claim gemacht wird basierend auf Testergebnissen, dieser Claim aber nicht valide ist. z.B. weil Besonderheiten des tatsächlichen Anwendungskontexts	▪ Methodensammlung und Hilfestellung bei der Auswahl z.B. Statistisches Testen, klare Metriken auf der rechten Seite des V-Modells, NN-Stresstests mit Adversarial Testing, NN-Analyse mit heat-maps etc. ▪ Konstruktive Hilfestellung bezüglich des Aufbaus der Sicherheitsargumentation

	geeignete Tangibles (Testberichte, Analysen etc.) als Evidenzen nutzt.

f) Während der Post-Market Surveillance

Problem	Lösungsansätze der VDE-AR-E 2842-61
Die Assurance Cases enthalten viele Annahmen zum Anwendungskontext. Diese können sich als in der Praxis nicht zutreffend erweisen.	▪ Das System wird nur für die den Einsatz gemäß der Zweckbestimmung (inkl. Nutzungsumgebung) „zugelassen“ ▪ Den Tatsächlichen Einsatz des Systems im Rahmen der Post-Market-Surveillance beobachten und fortlaufend prüfen, ob die Annahmen der Assurance Cases erfüllt sind ▪ Schrittweises Einführen der Systeme ggf. unter Beobachtung („Bootstrap Ansatz“)

4. Die Normenfamilie VDE-AR-E 2842-61

a) Anwendungsbereich

Die VDE-AR-E 2842-61 ist für alle Branchen und alle Applikationen anwendbar, die zur Klasse der autonom kognitiven Systeme, insbesondere der KI-Systeme zählen. Sie bezeichnet diese Systeme auch als „System of Systems“. Dabei wären unter „Systems“ im Sinne des Medizinprodukterechts eher „Produkte“ zu verstehen, keine Systeme im Sinne des Artikels 22 („Systeme und Behandlungseinheiten“).

Die VDE-AR-E 2842-61 hat aber keinen spezifischen Bezug zu Medizinprodukten.

Dennoch empfiehlt sich die Norm auch bei Medizinprodukten für die Sicherheitsargumentation mit Behörden und Benannten Stellen. Sie fügt bestehenden Regularien neue Aspekte wie die **Uncertainty** hinzu.

Sie VDE-AR-E 2842-61 empfiehlt sich

- bei der Entwicklung neuer Produkte / Systeme,
- bei der Weiterentwicklung dieser Produkte / Systeme und

VDE-AR-E 2842-61-2	Management	verfügbar
VDE-AR-E 2842-61-3	Development at Solution Level	abgeschlossen, in Freigabe
VDE-AR-E 2842-61-4	Development at System Level	erwartet für 2021-Q2
VDE-AR-E 2842-61-5	Development at Technology Level	erwartet für 2021-Q2
VDE-AR-E 2842-61-6	After Release of the Solution	verfügbar
VDE-AR-E 2842-61-7	Application Guide	zurückgestellt

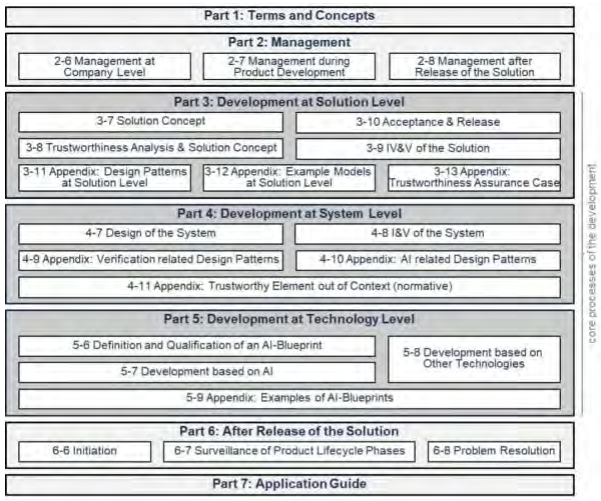

Abbildung 4: Übersicht über die Familie der Normen der VDE-AR-E 2842-61 (zum Vergrößern klicken) Quelle: Figure 2 der VDE-AR-E 2842-61-1

c) Struktur und Beispiele

Konzept der Norm	Beispiel

Für jedes Kapitel verlangt sie von den Verantwortlichen („Responsible“), Ziele festzulegen.	▪ Relevante Aspekte der Trustworthiness bestimmen ▪ Gefährdungen identifizieren und Risiken abschätzen ▪ Trustworthiness Goals festzulegen (d.h. das Ziel besteht darin andere Ziele zu bestimmen) ▪ „Trustworthiness measures“ zu bestimmen und dem „Solution Konzept“ zuzuordnen
Um das jeweilige Ziel zu erreichen, müssen bestimmte Aufgaben („Task“) erledigen werden.	Das Ziel „Relevante Aspekte der Trustworthiness bestimmen“ setzt die Norm mit der gleichlautenden Aufgabe gleich.
Dabei besteht jede Aufgabe aus einem Satz an Aktivitäten („Activity“).	Zu den 10 Aktivitäten zählen das Identifizieren von relevanten Normen (z.B. IEC 61508) und das Festlegen der „Trustworthiness Aspects“ (z.B. die der ISO 25020).
Für einige dieser Aktivitäten legt die Norm die zu berücksichtigenden Inputs fest.	Um die Trustworthiness Aspects festzulegen, muss der/die Verantwortliche die User Requirements und die regulatorischen Anforderungen berücksichtigen.
Für diese Aktivitäten können Hilfsmittel („Mean“) wir Werkzeuge, Templates oder andere Ressourcen festgelegt werden.	Zu den Hilfsmitteln zählen beispielsweise die o.g. Normen sowie Literaturhinweise.

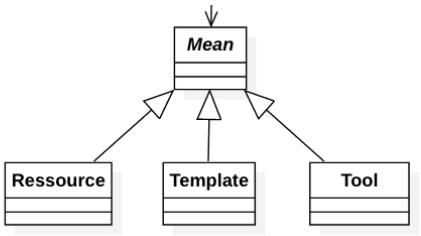

Abb. 5: Das Konzept der VDE-AR-E 2842-61 als UML-Klassendiagramm

5. Verbindlichkeit der VDE-AR-E 2842-61

Die VDE-AR-E 2842-61 ist keine harmonisierte Norm. Sie ist auch nicht für die Harmonisierung vorgesehen. Die Wahrscheinlichkeit, dass ein Auditor oder Prüfer einer Benannten Stelle diese Normenfamilie als Stand der Technik einfordert, ist (noch) gering.

Die Konzepte der Normenfamilie ergänzen sich gut mit denen der Normen ISO 14971, IEC 60601 und IEC 62304. Das gilt insbesondere für

- das Risikomanagement,
- den „risk based approach,
- das Dokumentations- bzw. Entwicklungs- und Lebenszyklusmodell (das ans V-Modell angelehnt erscheint).

6. Zusammenfassung

a) Was gefällt

Die Normenfamilie VDE-AR-E 2842-61 geht sehr systematisch vor. Sie verwendet ein Datenmodell und nutzt selbst eine klare und weitgehend umfassende Terminologie.

Es gefällt auch, dass sie den ganzen Lebenszyklus autonom kognitiver Systeme wie KI-Systeme abdeckt und ihren Aufbau an diesen Lebenszyklusphasen ausrichtet. Das erleichtert die Zuordnung.

Die Autoren sind offensichtlich in logischen Strukturen und Konzepten denkende Expert:innen für KI-Systeme.

b) Was nachdenken lässt

> The person responsible for the solution level shall use the knowledge gained on hazards and de- graded modes to define trustworthiness measures to cover all trustworthiness goals and further constraints given by their attributes (e.g. timely detection and control of relevant functional – consider safe state).
>
> Quelle: VDE-AR-E 2842-61-3, Kapitel 8

Zwar enthalten die Anhänge Beispiele. Dennoch wäre es wünschenswert gewesen, wenn nicht ausgerechnet der siebte Teil („Application Guide“) zurückgestellt worden wäre.

Wer wie beispielweise McKinsey-Berater das MECE-Prinzip und das „Pyramid Konzept“ verinnerlicht hat, wird sich fragen, ob die Hierarchie der Konzepte ausreichend trennscharf ist. Das folgende Beispiel stammt aus dem achten Kapitel des dritten Teils (Section 3-8):

Element	Beispiel aus der VDE-AR-E 2842-61	Kommentar
Ziele (Objectives)	The objectives of this section are: (1) to define the applicable trustworthiness aspects and to integrate relevant analysis methods from other standards;	Das ist eher eine Aufgabe al sein Ziel. Das eigentliche Ziel dieser Sektion besteht darin ein „Trustworthy Solution Concept“ zu entwickeln.
Aufgaben (Tasks)	to define the applicable trustworthiness aspects and to integrate relevant analysis meth- ods from other standards;	Das ist wortgleich der Formulierung eines der Ziele.
Aktivitäten (Activities)	The person responsible for the solution level shall define the applicable trustworthiness aspects.	Diese Formulierung gleich nahezu der vorangegangenen.

c) Fazit

Medizinprodukte keine autonom kognitiver Systeme sind.

Die Anwender der VDE-AR-E 2842-61 müssen in der Lage sein, in abstrakten Konzepten zu denken und die darin genannten Best Practices auf den konkreten Anwendungsfall anzuwenden. Das setzt hohe Kompetenzen voraus. Kompetenzen, die man von Personen erwarten muss, welche KI-Systeme für die Medizin entwickeln.

Die VDE-AR-E-2842-61 ist beim VDE verfügbar.

An diesem Artikel haben Dr. Rasmus Adler vom Fraunhofer IESE und Dr. Henrik Putzer von der fortiss, dem Landesforschungsinstitut des Freistaats Bayern für softwareintensive Systeme, und der Cogitron mitgewirkt. Beide stehen bei Rückfragen gerne zur Verfügung.

- rasmus.adler@iese.fraunhofer.de – Program manager „autonomous systems“ Fraunhofer IESE
- putzer@fortiss.org – Kompetenzfeldleiter Truistworthy Autonomous Systems bei der fortiss GmbH (fortiss.org)
- henrik.putzer@cogitron.de – Geschäftsführer und Berater bei der cogitron GmbH (cogitron.de)

Wie hilfreich war dieser Beitrag?

Bitte bewerten Sie:

Durchschnittliche Bewertung 3.8 / 5. Anzahl Bewertungen: 10

Category: Systems Engineering bei Medizinprodukten Von Prof. Dr. Christian Johner
13. April 2021 2 Kommentare

Schlagwörter: IEC 60601-1: Medizinische elektrische Geräte | Medizinische Software | Software-Architektur (IEC-62304-konform)

Teilen Sie diesen Post