Technische Aspekte der neuen Anwendungsregel VDE-AR-E 2842-61

Entwicklung und Vertrauenswürdigkeit von autonom/kognitiven Systemen

fortiss GmbH
cogitron GmbH

2021-03-16

VDE DKE Tagung Funktionale Sicherheit (online)

Gestern im KI-Labor ...

Und damit machst Du **maschinelles Lernen**?

Klar! Man kippt die Daten in diesen Haufen voller lineare Algebra, und dann bekommt man die Antworten hinten raus.

Und was, wenn die Antworten falsch sind?

Einfach den Haufen umrühren bis es richtig aussieht.

KI liefert (teils) erstaunliche Erfolge

Es bleiben offene Fragen!

Nachweisliche Vertrauenswürdigkeit

Standortbestimmung
Künstliche Intelligenz ist hochaktuell

Quelle: www. gartner.com

Gartner hype-cycle

Die Bundesregierung

Strategie Künstliche Intelligenz

KI als Markenzeichen für Deutschland

"KI made in Germany" soll zu einem internationalen Markenzeichen für moderne, sichere und gemeinwohlorientierte KI-Anwendungen auf Basis des europäischen Wertekanons werden. Damit das gelingt, hat das Kabinett die von BMWi, BMBF und BMAS gemeinsam vorgelegte Strategie Künstliche Intelligenz beschlossen.

12. September 2020, 11:02 Uhr Software

Die meisten künstlichen Intelligenzen sind strunzdumm

As of A

Künstliche Intelligenz ist bereits auf dem Markt (Risiken sind real)

Quelle: www.npr.org

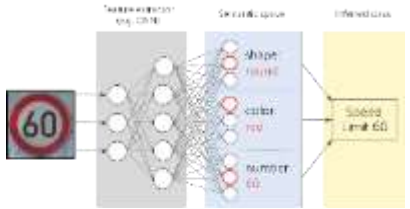
Quelle: www.amazon.de (Alexa)

Quelle: www.LG.com (LG V40 ThinQ) mobileye.de

Quelle: www. mobileye.de

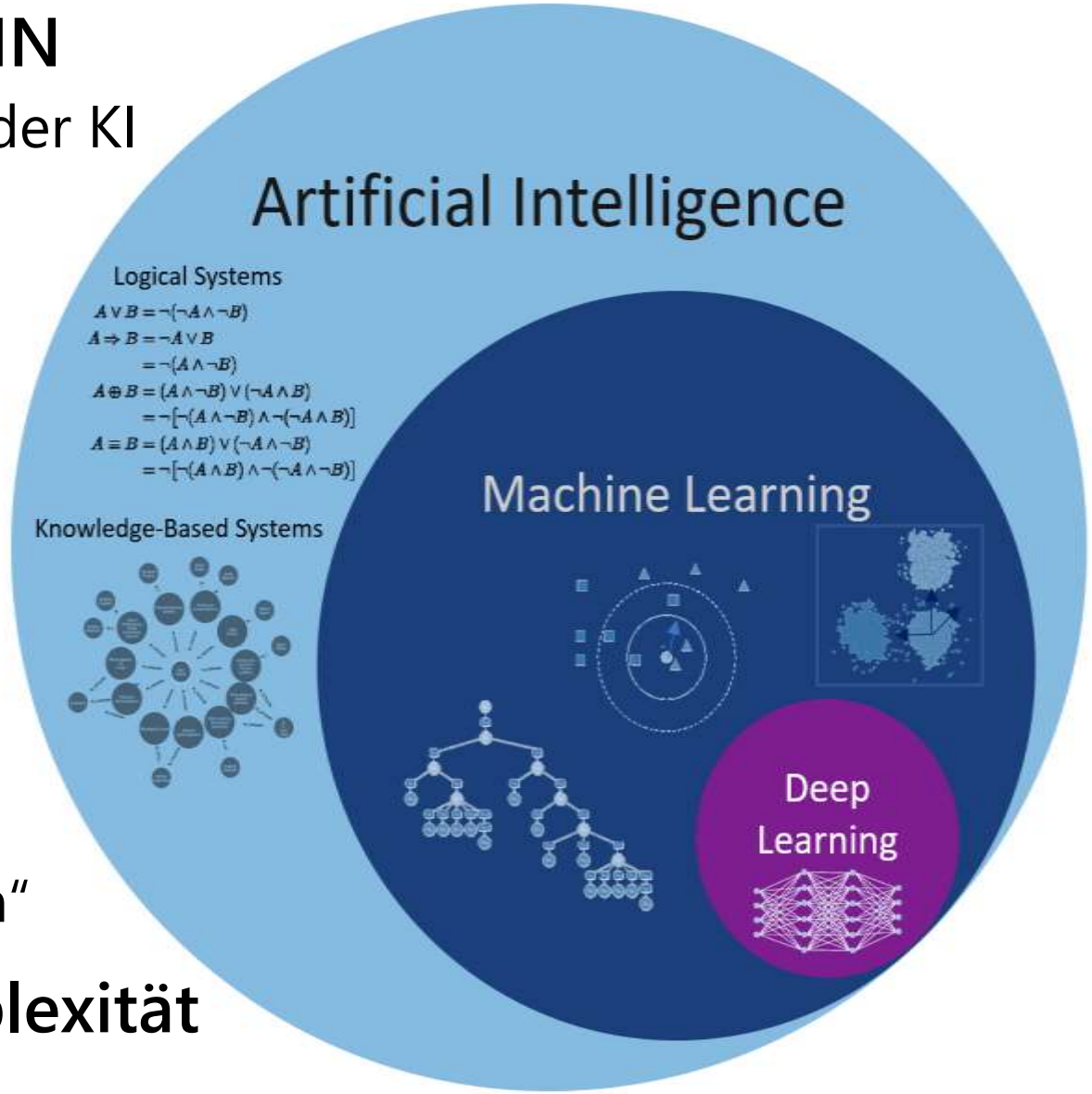
- **KI besiegt weltbesten Go-Spieler**
  - (Lee Sedol vs. DeepMind)
- **KI besiegt Profi-Spieler in Starcraft II**
  - (DeepMind)
- **Ausassistent erledigt Aufgaben auf Zuruf**
- **KI Verbessert Handy-Bilder**
- **Assistenten erkennen Verkehrsschilder und kritische Situationen** (Fußgänger)
- **„schwache KI“ (Produkte) vs. „starke KI“ (ggf. Forschung)**
- **HW, SW, KI ≠ Risiken**
  - Fehlerhaftes Design
    - Human Factors
    - Handwerkliche Fehler
  - Fehlerhafte Kommunikation
  - Böswillige Nutzung

Forschung: www.fortiss.org

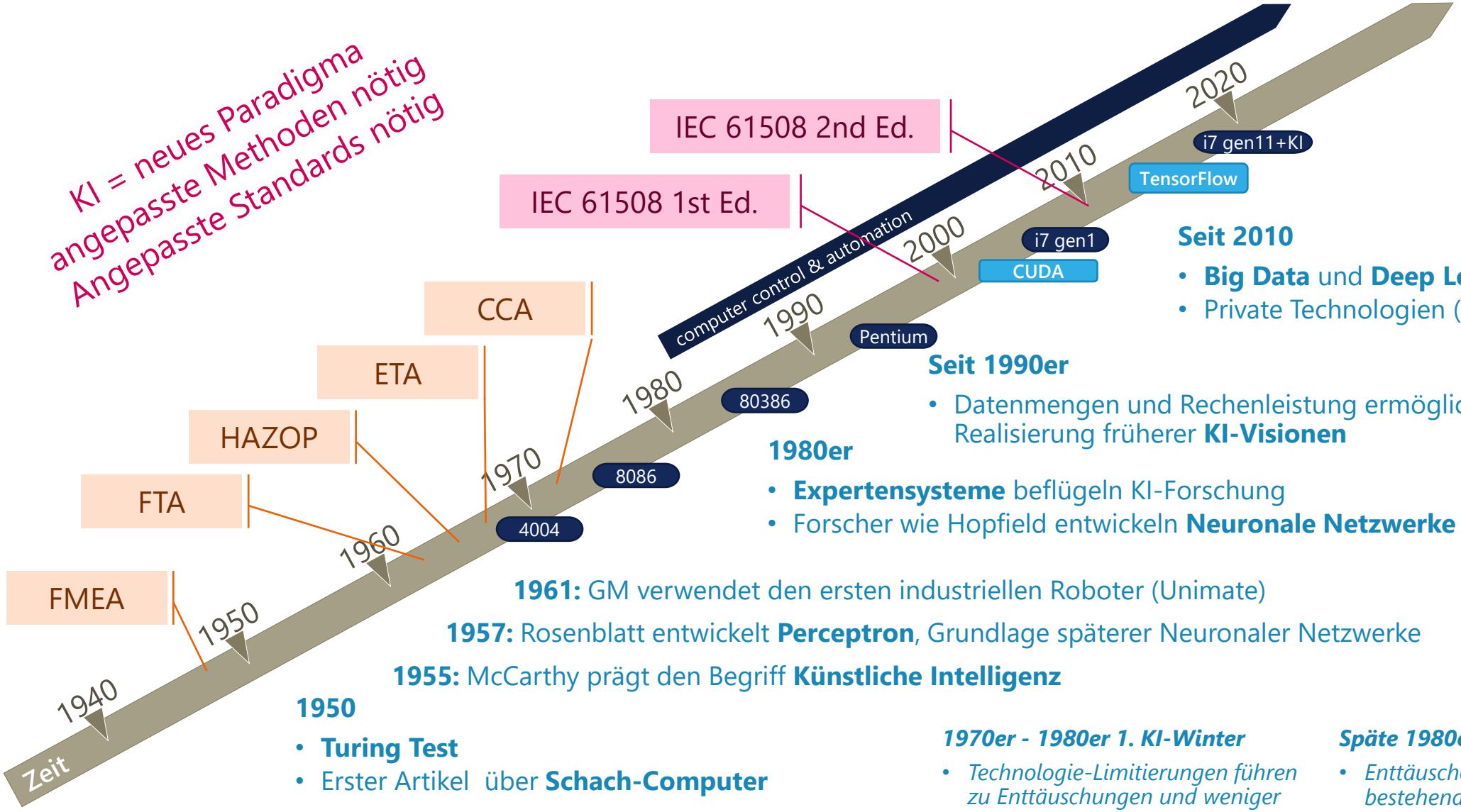


Abgrenzung: KI ~ ML ~ DL/NN

- **Keine Synonyme: KI ~ ML ~ DL ~ NN**
  - Maschinelles Lernen ist ein Teilbereich der KI
    - Symbolisch
    - Sub-symbolisch (konnektionistisch)
  - Neuronales Netz
    ist eine Ausprägung von ML
- **Eigenschaften von NN (DL)**
  - Realisieren implizite Anforderungen
  - Klassischen Verifikation unmöglich
  - Sind deterministisch
  - Lernen nicht automatisch immer weiter
  - NN kann nicht „verstehen“ oder „wollen“
- **Ermöglichen hohe Funktions-Komplexität**



Zeitlinie von Funktionaler Sicherheit und Künstlicher Intelligenz



Seit 2010

- **Big Data** und **Deep Learning**
- Private Technologien (digitale Assistenten)

Seit 1990er

- Datenmengen und Rechenleistung ermöglichen die Realisierung früherer **KI-Visionen**

1980er

- **Expertensysteme** beflügeln KI-Forschung
- Forscher wie Hopfield entwickeln **Neuronale Netzwerke** weiter

1961: GM verwendet den ersten industriellen Roboter (Unimate)

1957: Rosenblatt entwickelt **Perceptron**, Grundlage späterer Neuronaler Netzwerke

1955: McCarthy prägt den Begriff **Künstliche Intelligenz**

1950

- **Turing Test**
- Erster Artikel über **Schach-Computer**

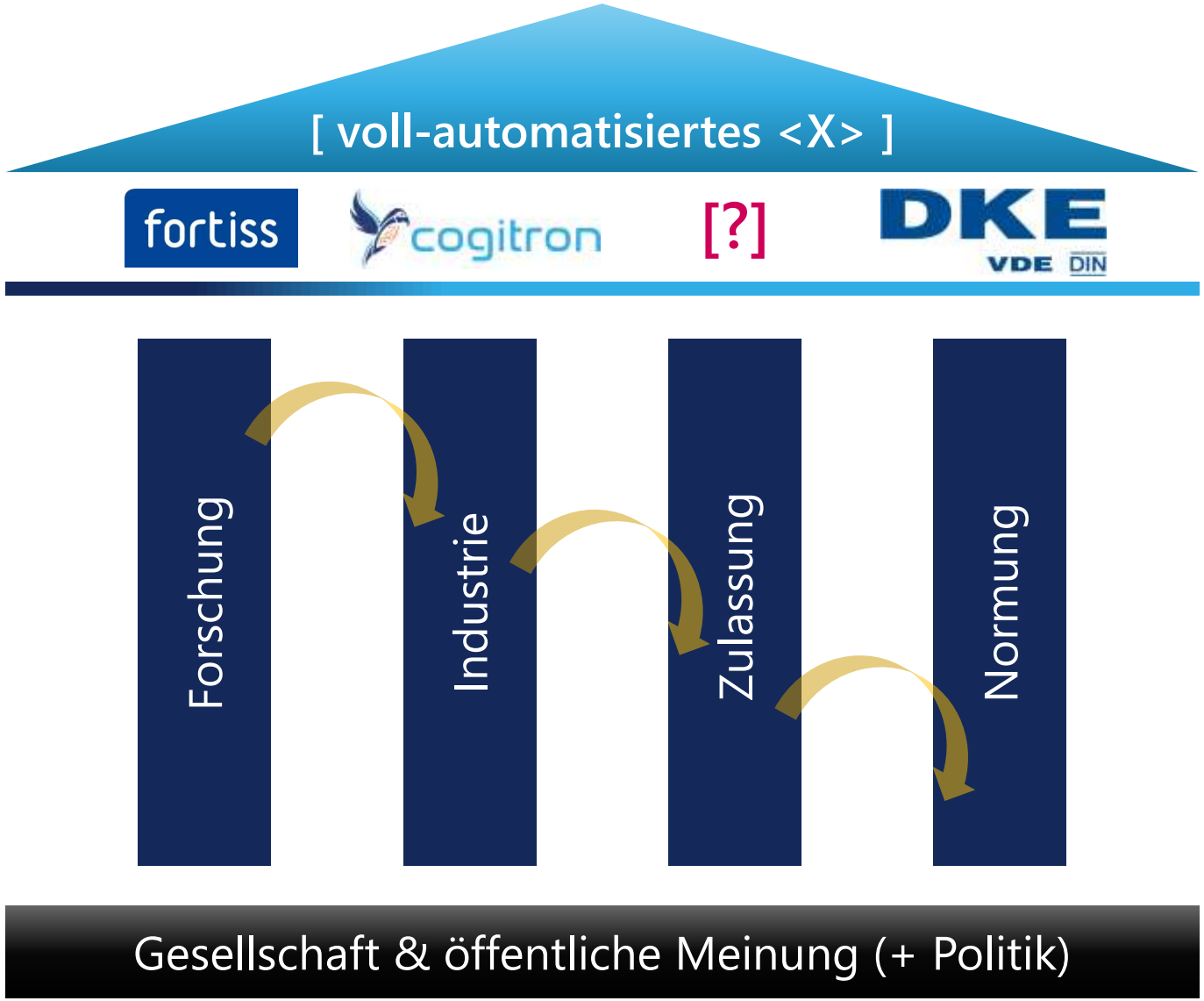
1970er - 1980er 1. KI-Winter

- *Technologie-Limitierungen führen zu Enttäuschungen und weniger Forschung*

Späte 1980er - 1990er 2. KI-Winter

- *Enttäuschende Ergebnisse bestehender KI-Systeme*

Inkubation der neuen Technologie



- **Gesellschaft & Politik**
  - Mythen statt Verständnis
  - Ringt um Regularien

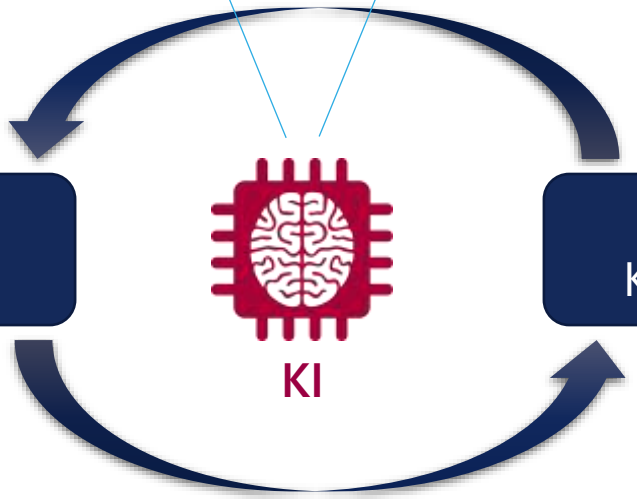
Akzeptanz & Ethik ?

- **Forschung**
  - Präsentiert erstaunliche Anwendungen
  - Technologie noch nicht unter Kontrolle
- **Industrie**
  - Teilweise weiter als Forschung (massive finanzielle Mittel)
  - Zwischen Unwissenheit & Überschätzung
  - Fehlende Fundierung / Standards
- **Zulassung**
  - Keine klaren Grundlagen/Gesetzte
  - Zulassung aktuell nicht handhabbar (Teufelskreis Entwicklung & Zulassung)
- **Normung**
  - Unübersichtliche Normungslandschaft
  - Viele nebenläufige Gremien
  - Erste Roadmaps & Standards verfügbar
    - Deutsche Normungs-Roadmap (auch andere)
    - VDE-AR-E 2842-61

Teufelskreis um neue Technologien (u.a. KI)

game changer
Artificial Intelligence

Versprechen: neue Funktionen &
Komplexitätsbeherrschung



VDE-AR-E 2842-61

- **Wie realisiert man Produkte mit KI?**
- **Wie verifiziert man KI (mit Sicherheitsrelevanz)?**

nächste Schritte!

- **Wie lässt man KI-basierte Produkte zu?**
- **Welche Auswirkungen hat KI auf Unternehmen (und Märkte)?**

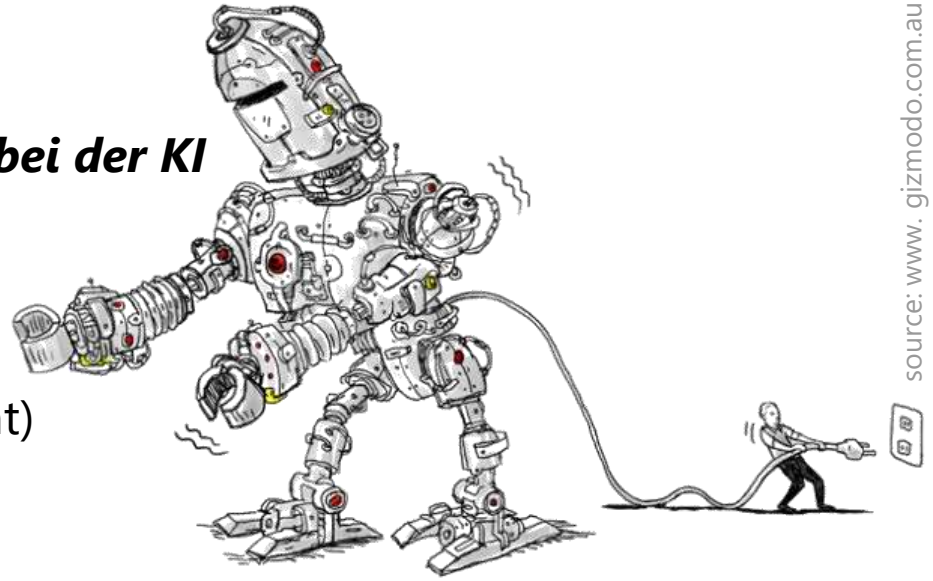
unsere Beobachtung:
Deutschland
hätte die Mittel
konkurrenzfähig zu sein

Stand der Technik – eine unaufgeregte Sichtweise

- **KI/ML/NN = Engineering-Methode / neue Technologie**
  - NN realisiert implizite Anforderungen aus Daten
  - NN-Technologie & -Eigenschaften nicht vollständig verstanden

  } *realisieren Automation (Regelfunktion)*

- **NN zeigen bekannte Probleme der Automation**
  - complexity / brittleness / opacity / literalism (Billings, 1997)
  - supervisory control (Sheridan, 1976)
  - **Verantwortung** liegt beim Entwickler, Hersteller & Nutzer - ***nicht bei der KI***



source: www. gizmodo.com.au

- **Lösung: Autonom/Kognitive Systeme**
  - Neues Paradigma:
    komplexe Systeme, Trustworthiness, KI-Technologie (KI Element)
  - Ziele: **Implementierung, Verifikation, Nachweis**
    (Grundlage für Prüfung/Zulassung und Zertifizierung)
  - Ansatz: **Framework**, Prozess, Methoden, Maßnahmen & Werkzeuge

  - VDE-AR-E 2842-61 „Entwurf und Vertrauenswürdigkeit von autonom/kognitiven Systemen“

Autonomous/Cognitive System (A/C-System)

independent from industry & application

separate ethical aspects from technology

risk based approach along lifecycle

Performance und Trustworthiness

Solution Level

System Level

Technology Level (AI)

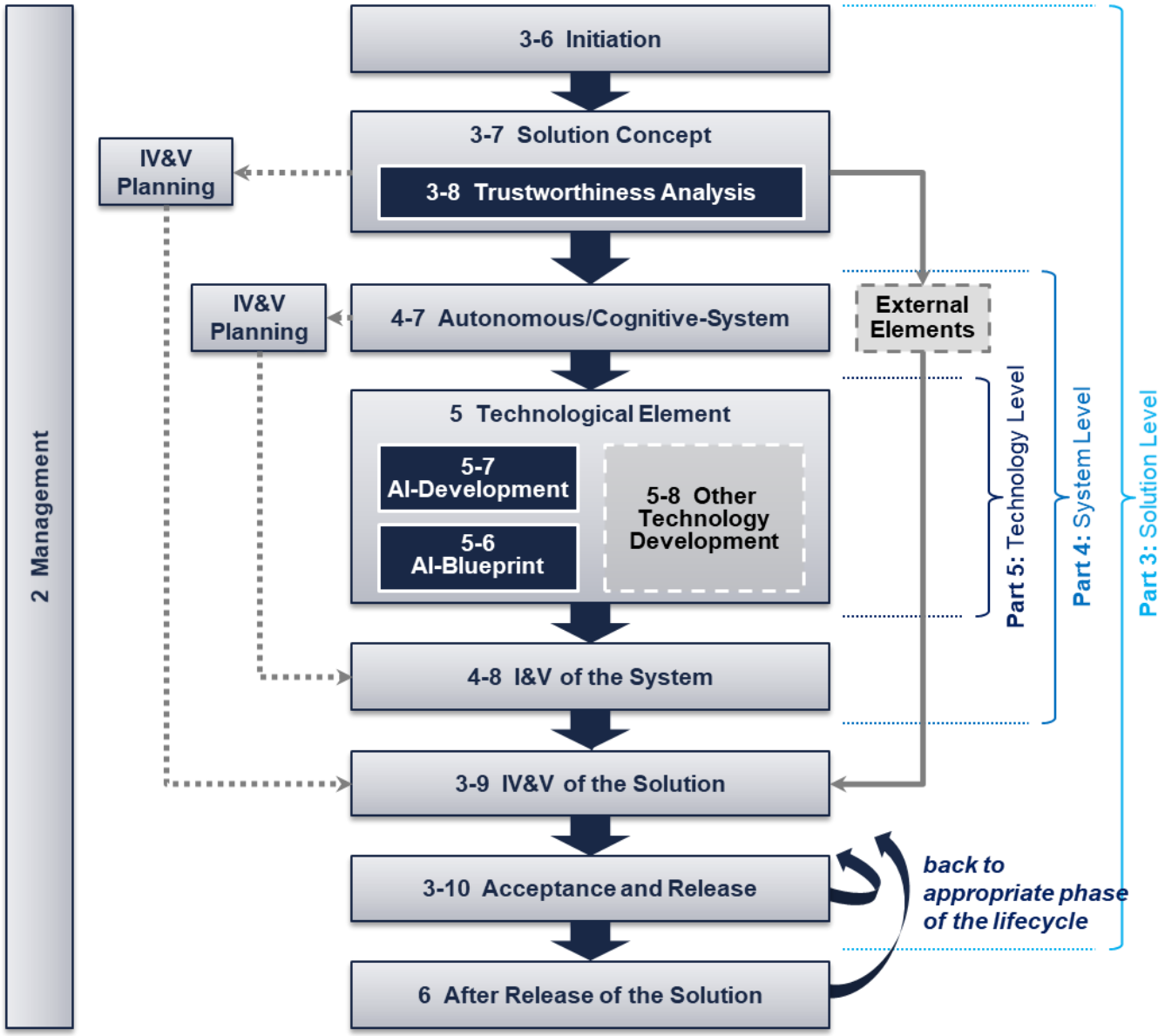
Safety
Security
Usability
Ethics
[...]

environment
other objects & systems
user & stakeholders

VDE-AR-E 2842-61

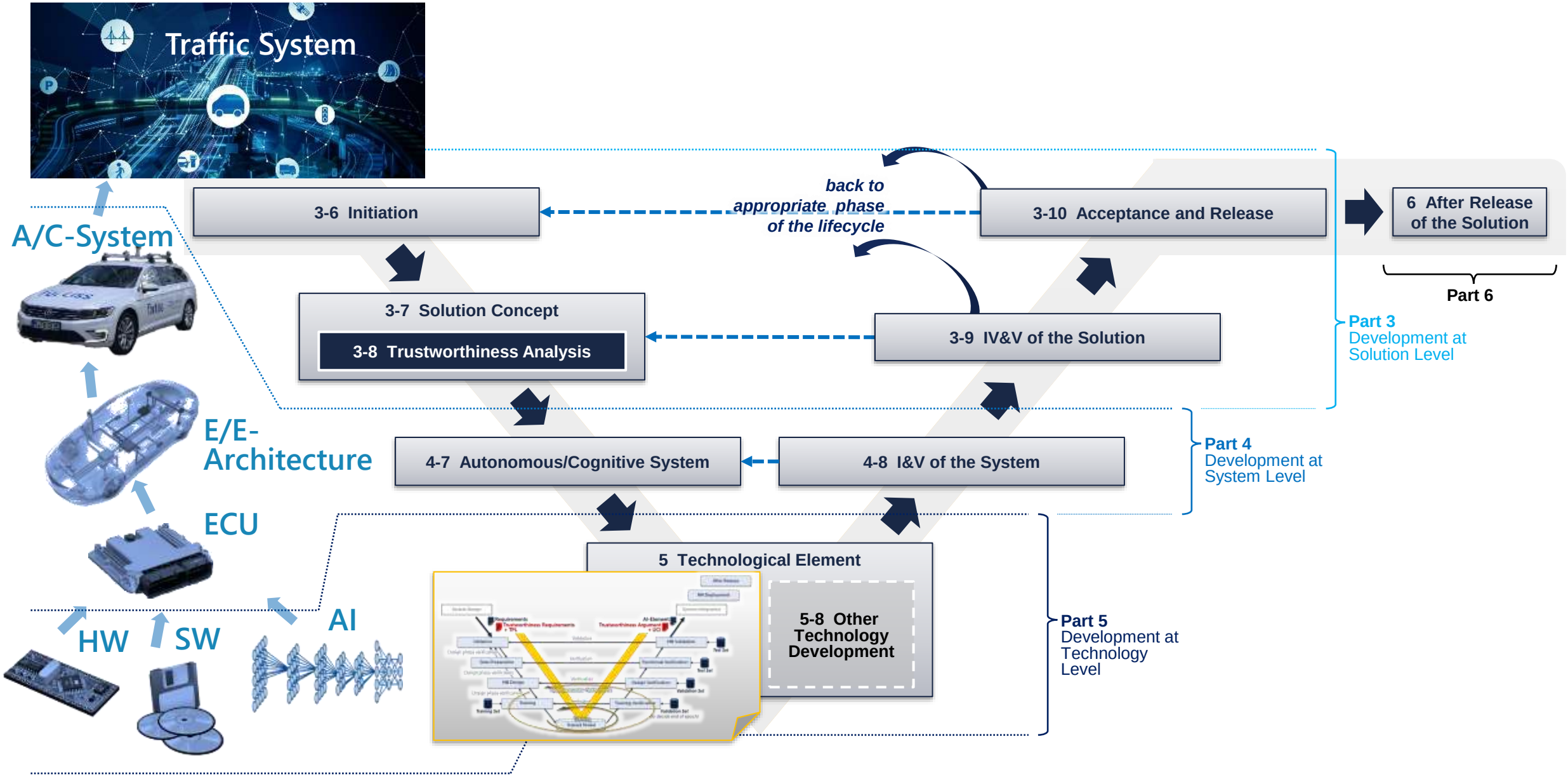
inspired by IEC 61508
(aware of other working groups, e.g. ISO/IEC JTC 1/ SC 42)

Der erweiterte Referenz-Lebenszyklus

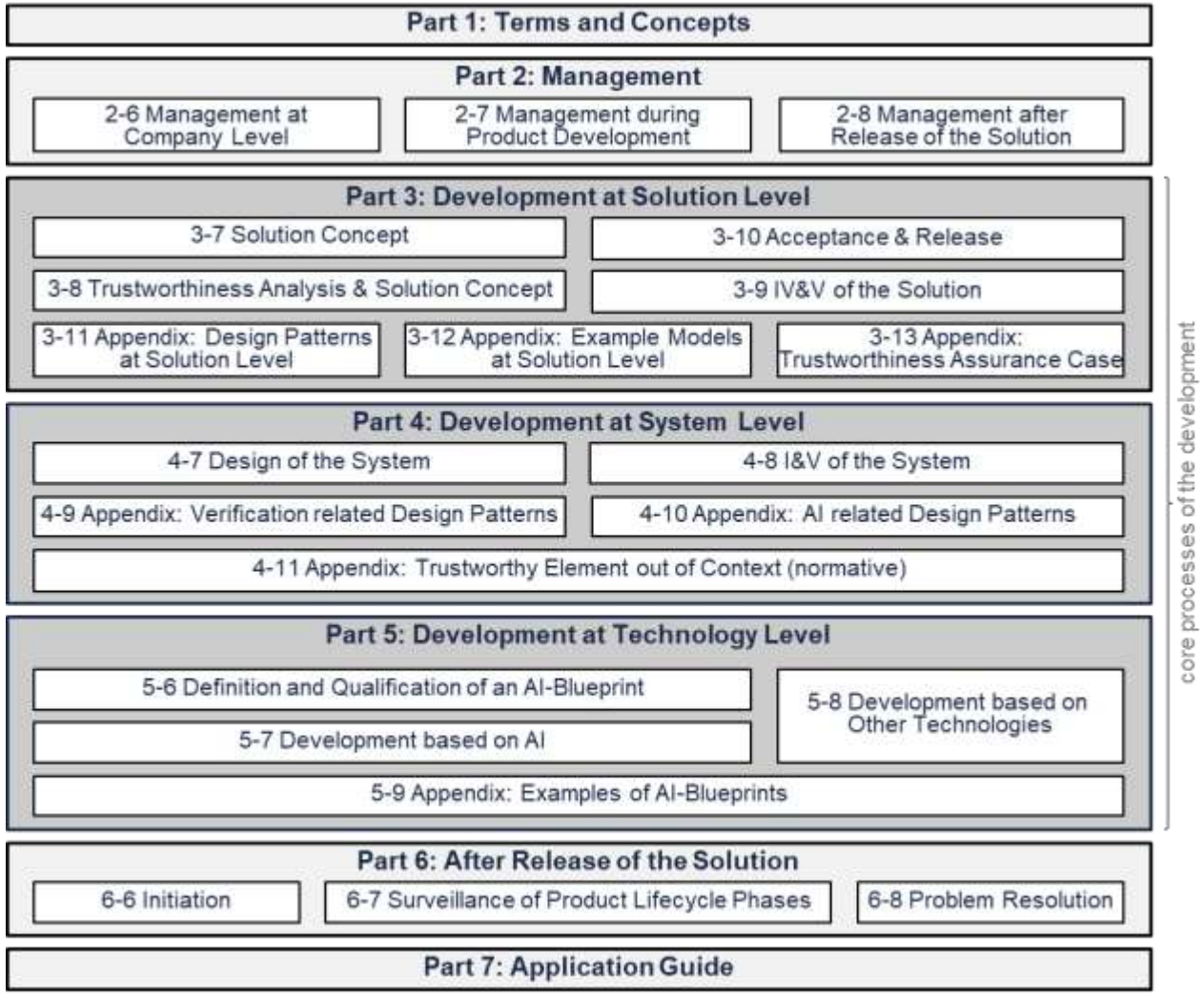


- **Strukturierter Ansatz**
  - Generischer Rahmen
  - Risiko-basiert
  - Referenz-Lebenszyklus (Phasen, Ziele, Aktivitäten, Artefakte)
  - Leicht integrierbar
- **Künstliche Intelligenz adressiert**
  - 3. Art der Technologie
  - KI-Blaupause (Prozess-Plugin spezifisch für eine KI-Technologie)
  - Neuer Ausfall-Typ
    - Systematisch
    - Zufällig
    - Unsicherheits-bezogen (UCI)
- **KI und mehr**
  - Solution Level
  - Trustworthiness Analysis & Trustworthiness Performance Level (TPL)
  - System Level (hierarchischer Ansatz)

Referenz-Lebenszyklus als V-Modell mit Abstraktionsebenen (Automotive)



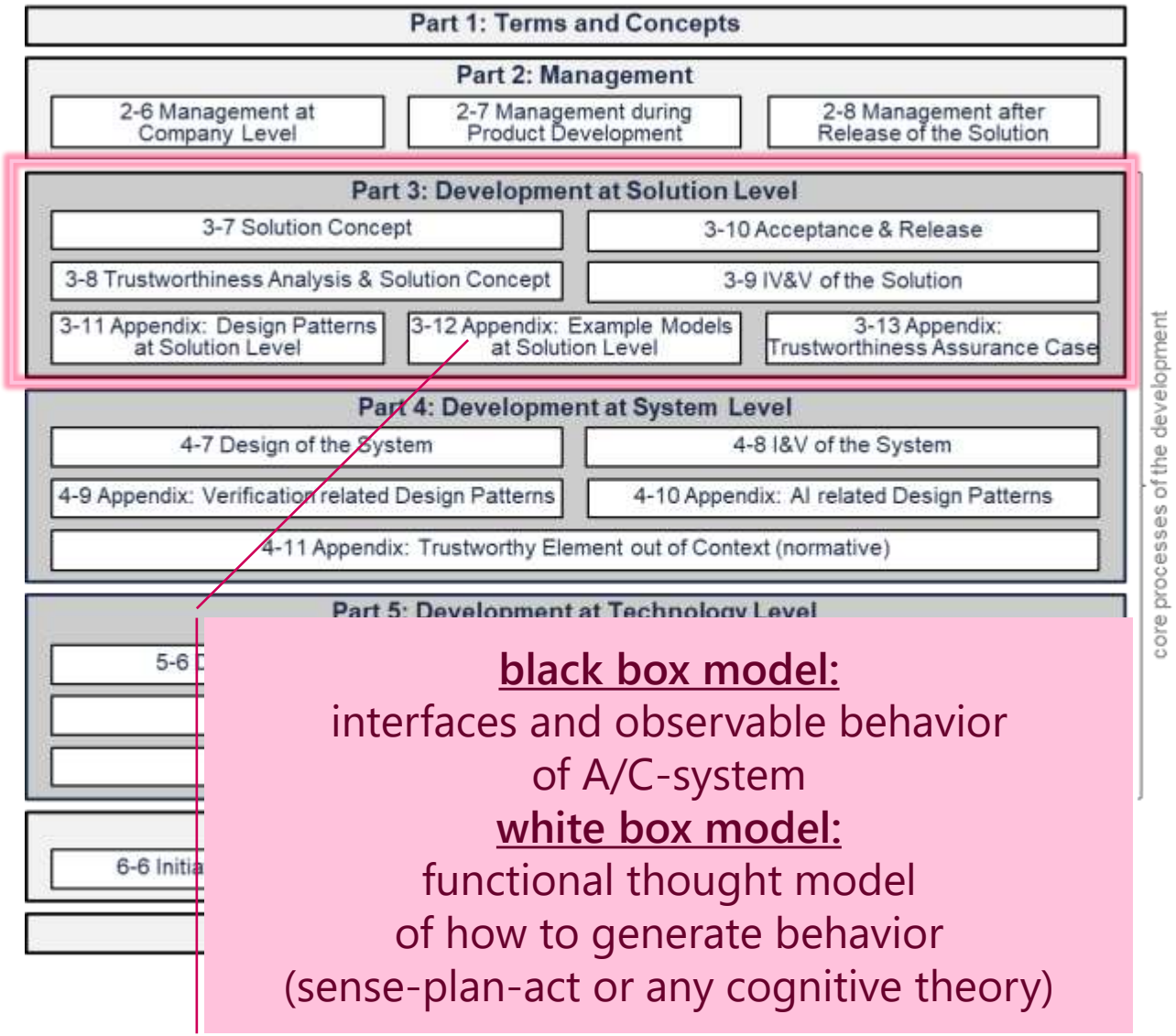
Grafische Übersicht



A framework for the development of trustworthy autonomous/cognitive systems captured in 6+X parts:

- part 1: Terms and Concepts ✓
- part 2: Management ✓
- part 3: Development at Solution Level ✓
- part 4: Development at System Level (soon)
- part 5: Development at Technology Level (soon)
- part 6: After Release of the Solution ✓
- part 7: Application Guide (postponed)

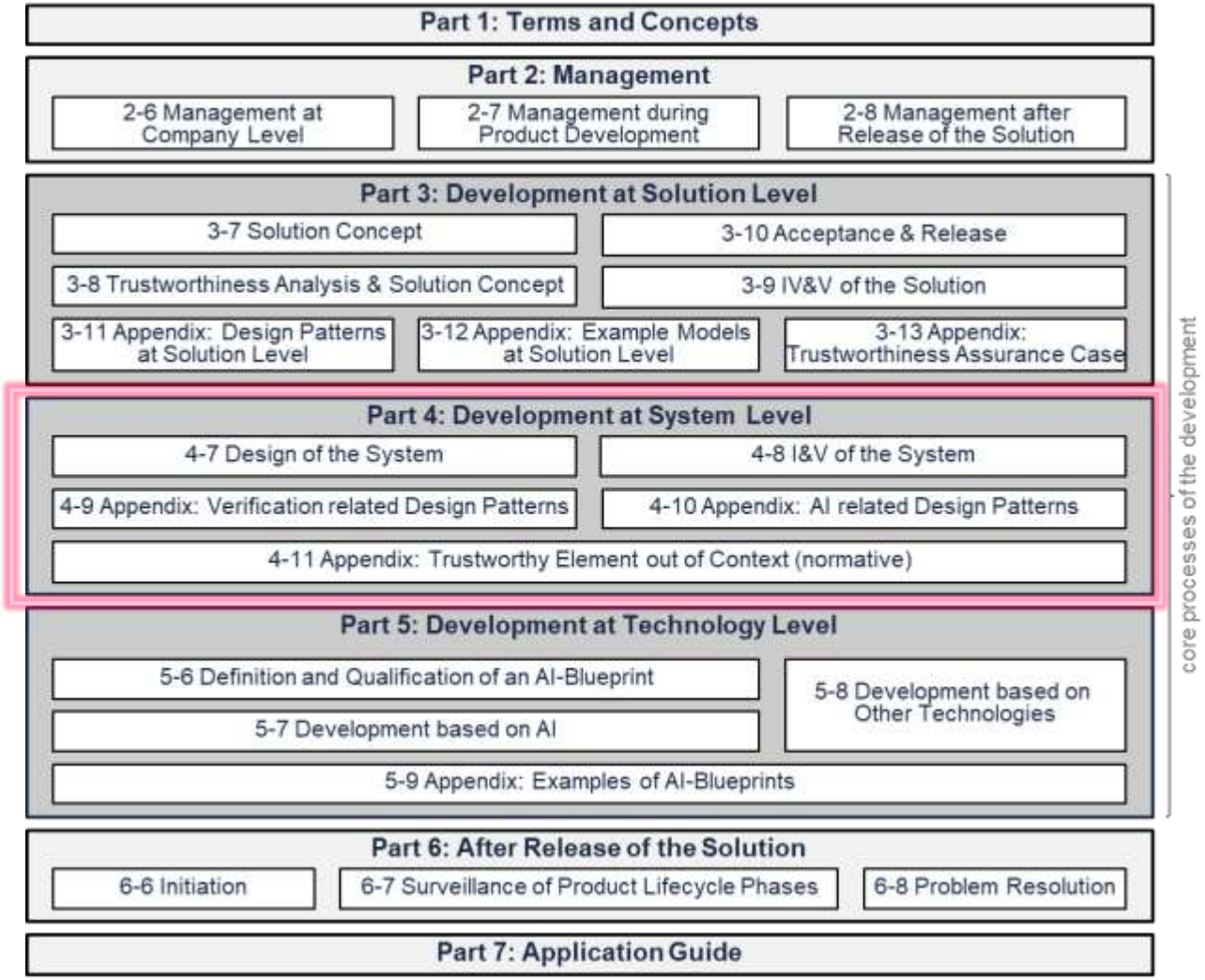
Teil 3: Development at Solution Level



black box model:
interfaces and observable behavior
of A/C-system
white box model:
functional thought model
of how to generate behavior
(sense-plan-act or any cognitive theory)

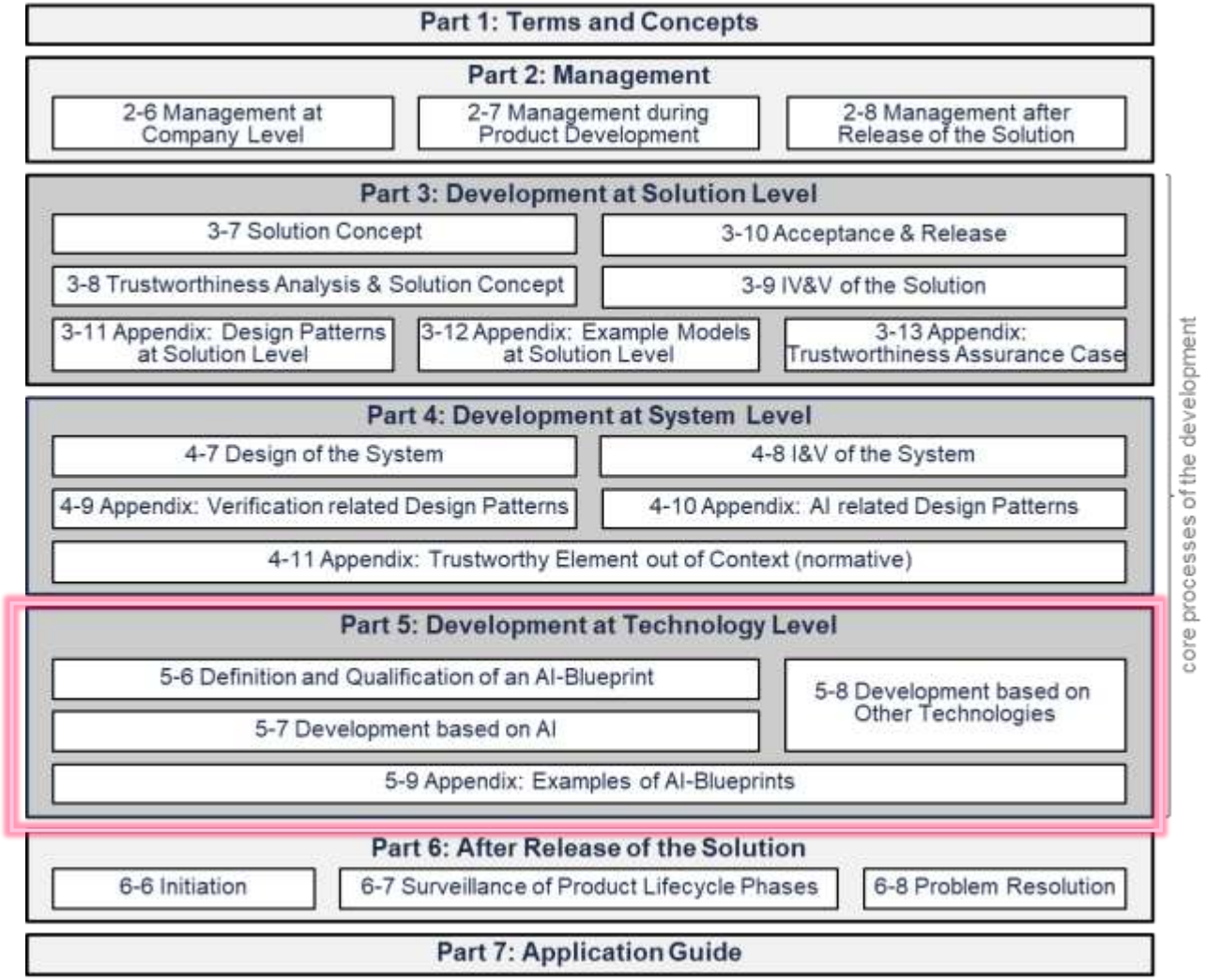
- **Initial requirements**
  - Use Cases (incl. environment)
  - Intended benefit
  - Prlim. Trustworthiness aspects
- **Environment as a System = Solution Concept**
  - Environment with all elements
  - A/C-system (AI might be a component)
  - User & all stakeholders
  - Ontology
- **Trustworthiness Analysis**
  - Use well known risk analysis mathods
  - Harmonize in
    - Trustworthiness goals (TPL)
    - Trustworthiness measures (TPL)
- **Trustworthiness solution concept**
- **Integration, Verification & Validation**
- **Acceptance & Release of the Solution**

Teil 4: Development at System Level



- **Hierarchical development of A/C-system**
  - Iterative application of part 4
  - Tracing of TPL and attributes (inheritance)
    - Allocation (vertical inheritance)
    - Decomposition (shared inharitance)
    - Independence (horizontal inheritance)
- **Design Patterns to support verification**
  - Monitors (incl. performant/safe switch over)
  - Functional redundance
  - Online verification
- **Design Patterns to ensure AI features**
  - Understand- | Explain- | Interpret-Ability
  - Continuous learning
  - Adaptive behavior (unknown unknowns)

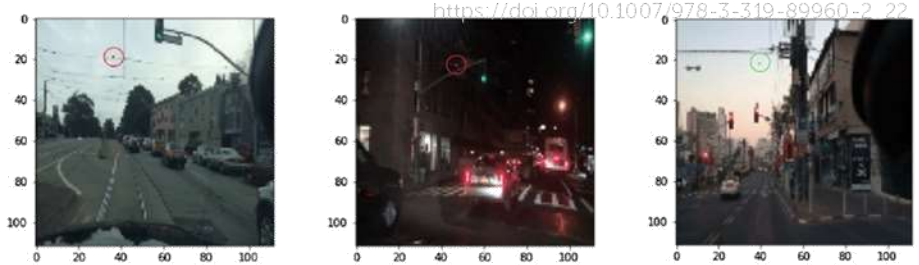
Teil 5: Development at Technology Level

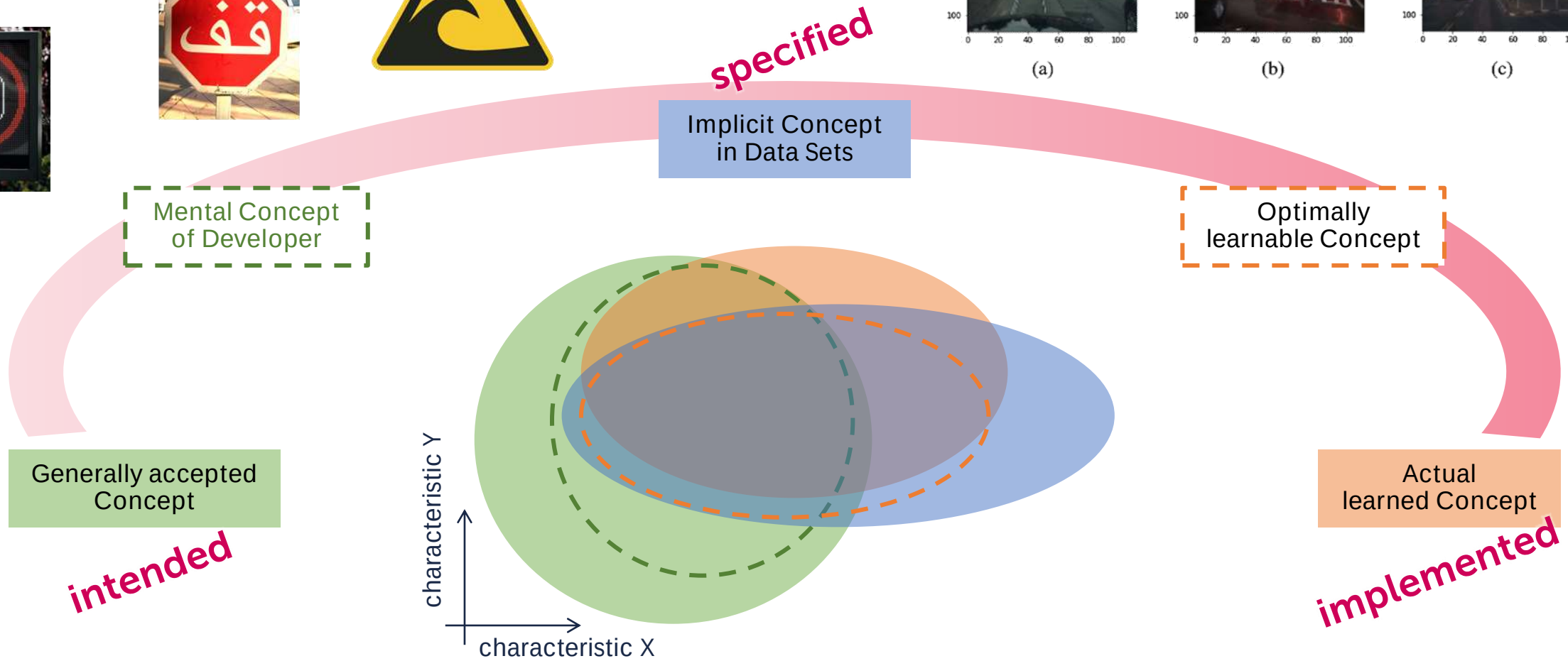


- **Implementation of Technology Elements**
  - Separate other technology from AI technology
  - Other technology handled through well proven standards (e.g. IEC 61508)
- **AI-blueprints – Definition & Qualification**
  - Ensure **plug & play interfaces** to lifecycle
    - Input interface
    - Output interface
  - Ensure **Usability**
    - Effectiveness: AI element can be built
    - Efficiency: development is not cumbersome
    - Easy to learn: e.g. known notation, min. overhead
    - Fun to use: loved by developers
  - Ensure **quality**
    - verification of AI-blueprint (e.g. review, PFMEA)
    - validation of AI-blueprint (incl. evaluation)
- **AI-blueprint application**

Probleme mit Neuronalen Netzen (Prozess)

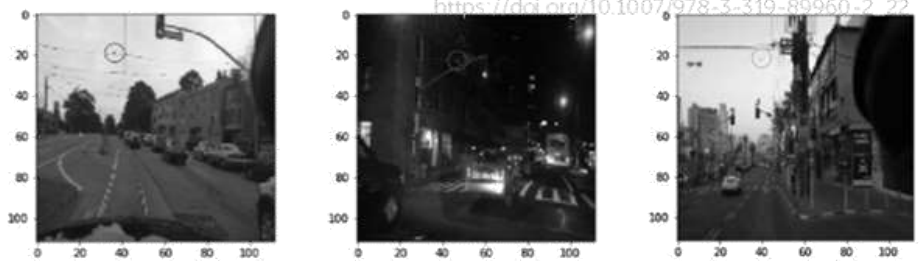
Source: D. Beyer and M. Huisman (Eds.): TACAS 2018, LNCS 10805, pp. 408–426, 2018.
https://doi.org/10.1007/978-3-319-89960-2_22





Probleme mit Neuronalen Netzen (Analyse)

Source: D. Beyer and M. Huisman (Eds.): TACAS 2018, LNCS 10805, pp. 408–426, 2018.
https://doi.org/10.1007/978-3-319-89960-2_22



specified

Implicit Concept

Specification

= *Pictures with Semantics by Labeling*

Data Set

Mental Concept of Developer

Optimally learnable Concept

rather systematic failures

failures due to characteristics of technology

Generally accepted Concept

intended

characteristic Y

characteristic X

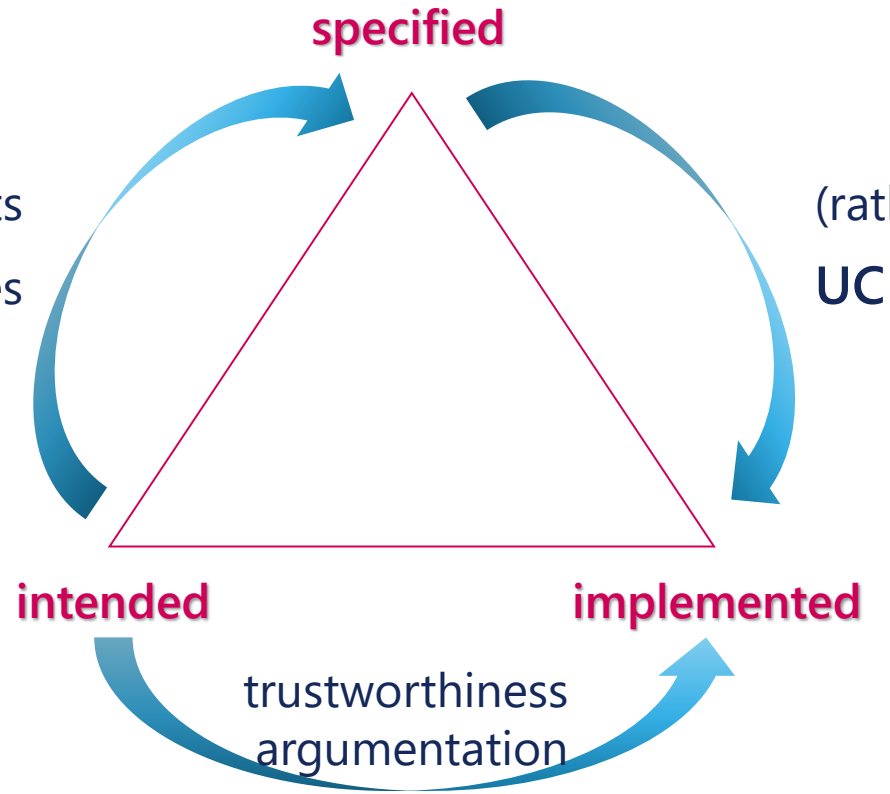
Actual learned Concept

implemented

Uncertainty-Confidence-Indicator (eine Art λ_{KI})

(rather) systematic faults

TPL: methods & measures



(rather) uncertainty-related faults

UCI: **u**ncertainty-related **c**onfidence **i**ndicator

KI dritte Art von Technologie mit einer dritten Art von Fehler

Anforderungen an die Plattform (HW & klassische SW)

type of failure	measures	measures for HW	measures for SW	measures for AI
systematic	Qualitative Requirements: Culture, Experts, QS Process, Design, Methods & Measures	systematic capability	systematic capability	systematic capability
random	Quantitative Requirements: Metrics and Thresholds	λ, SFF, DC, SIL-related target	-- / --	-- / --
uncertainty-related	Structured Approach: Metrics, References, Measures and Argumentation	-- / --	-- / --	Uncertainty confidence indicator (UCI)

"Trustworthiness Assurance Case" (z.B. GSN)

Metriken als Evidenzen in der Vertraulichkeits-Argumentation

Fragen zu Performanz, Unsicherheit und Determinismus

Wie gut muss das NN sein?
Wie gut sind Sie?
(Performanz/Zielwert)

- HUND
- MUFFIN

?

Wird ein gegebenes NN auf ein bestimmtes neues Bild reagieren?
~ Können wir die Klassifizierung vorhersagen?
(Unsicherheit)

Ein gegebenes NN wird den gleichen Input immer wieder gleich klassifizieren.
(Determinismus)

Probleme im Betrieb: Angriffe auf KI

◀ Neue Objekte

Missbrauch ▶

„Horizont“

fehl-Klassifizierung ▲

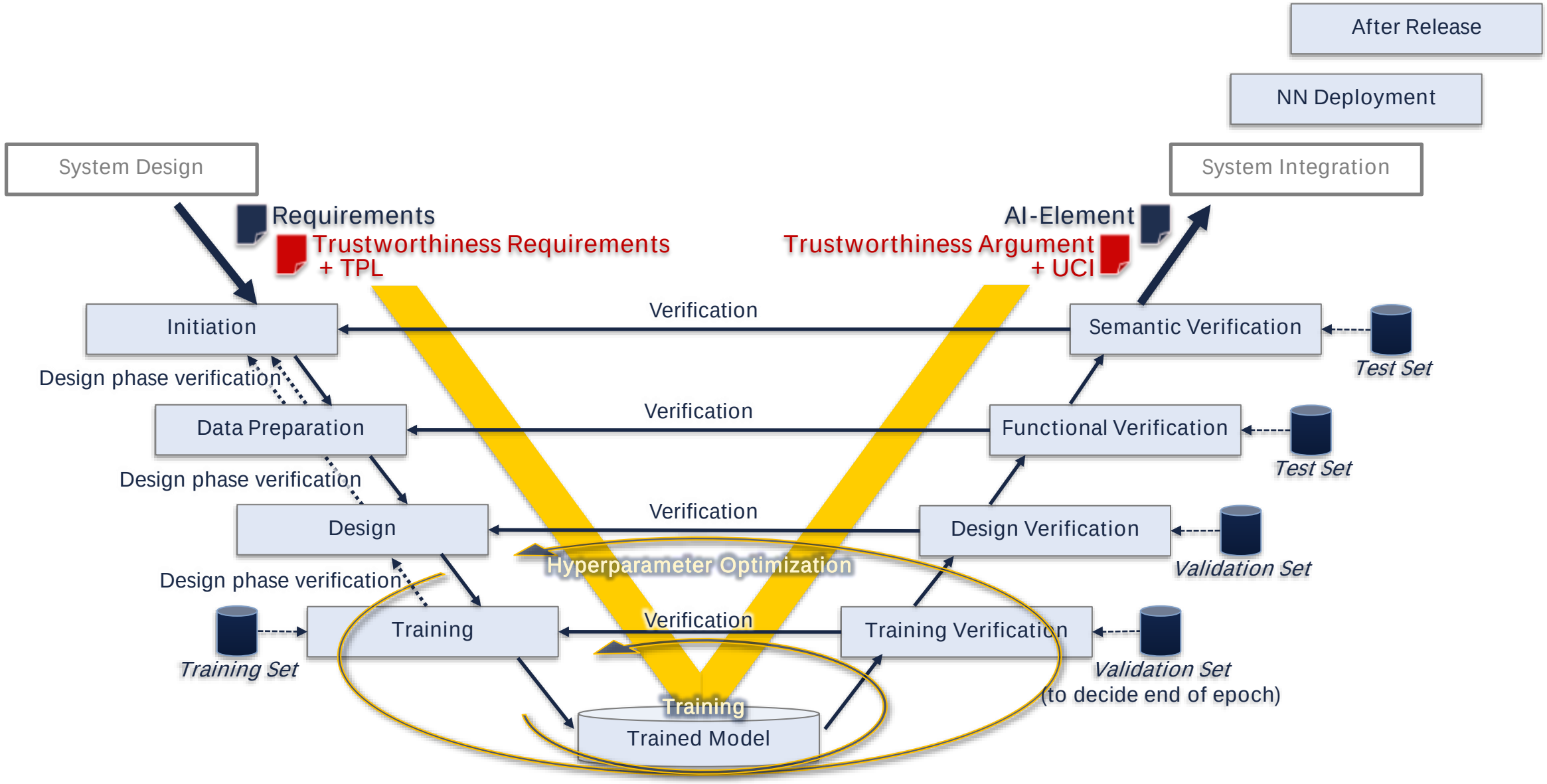
◀ Angriff
(3D street art)

Angriff ▶
(adversarial attack)

- **Probleme im Betrieb**
  - Missbrauch
  - Neue Objekte
  - Unvorhergesehen Situationen (edge cases)
  - Perzeptions-Attacken (3D-street-art)
  - Adversarial Patterns
- **Fragestellungen**
  - Wie spezifiziert man ein NN?
  - Wie entwickelt man ein NN strukturiert und nachvollziehbar?
  - Handwerkliche Fehler?
  - Wie „robust“ ist ein Neuronales Netz?

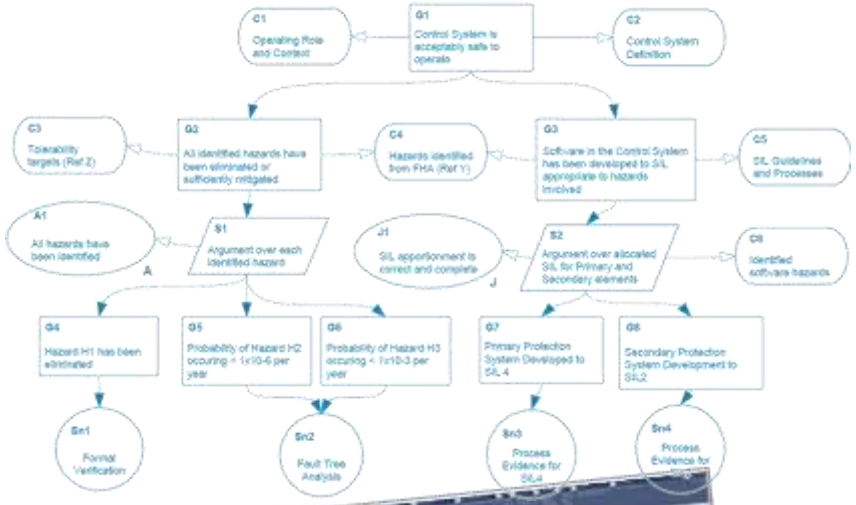
  - Wie sieht eine **KI-Blaupause** aus?

Beispielhafte KI-Blaupause (AI-blueprint) für tiefe Neuronale Netze



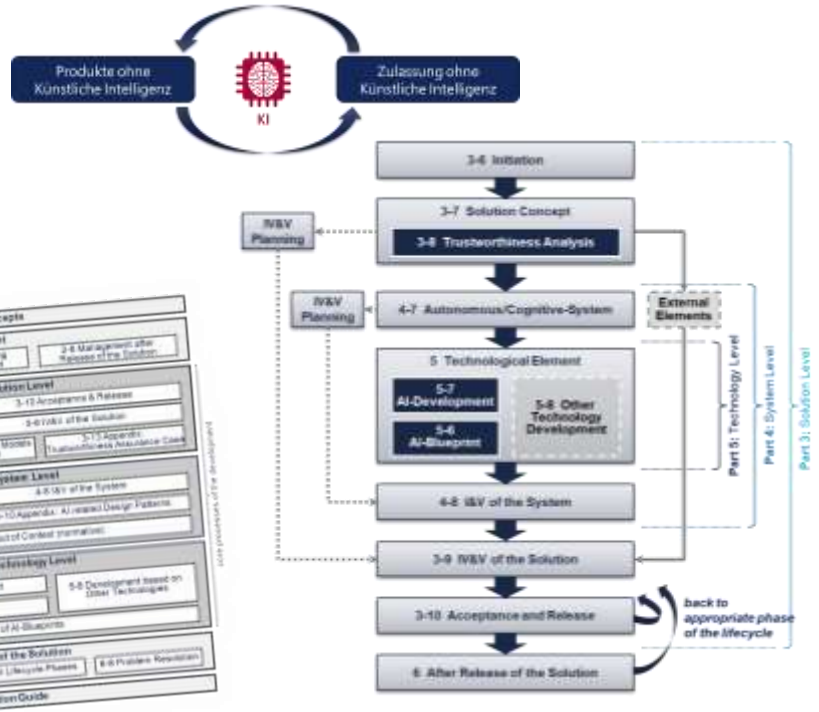
UCI Berechnungs-Ansatz: Quantitativer GSN mit NN-Metriken als Evidenzen (Ansatz in Entwicklung!)

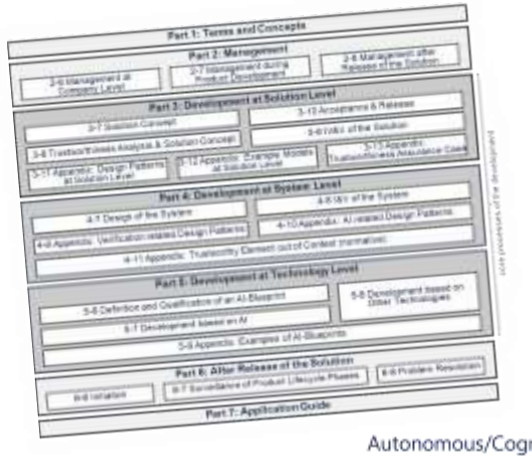
- **Sicherheitsargumentation Solution & System**
  - Argumentation vom TOP: „A/C-System is safe to be used in defined environment“
  - Argumentations-Muster (GSN-Design-Patterns)
- **Sicherheitsargumentation KI-Element (NN)**
  - Goal Structuring Notation
  - (Semi-) Quantitative Erweiterung (= Verallgemeinerung der quant. FTA)
  - Evidenzen (Basisevents) ergeben sich aus Metriken
- **Metriken für tiefe Neuronale Netze**
  - White Box, Black Box, Grey Box
  - Quelle: Forschung, Kataloge, etc.
  - KI-Absicherung: über 70 Metriken

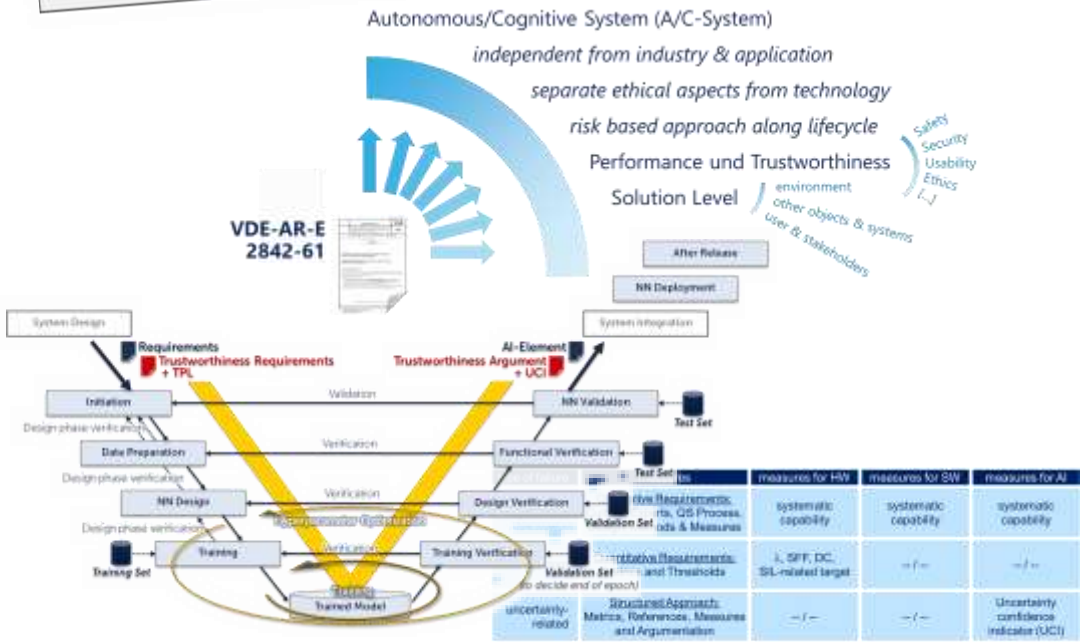


Zusammenfassung

- **VDE-AR-E 2842-61**
  Development and trustworthiness of autonomous/cognitive systems
  - Neue Anwendungsregel
  - Referenz für vertrauenswürdige KI
  - Risiko-basierter Ansatz
  - Unabhängig von Branche und Anwendung
- **KI berücksichtigt**
  - KI als Element im Systemdesign
  - KI als „dritte“ Technologie (neben HW+SW)
  - KI-Blaupause
- **Status & Ergebnisse**
  - 4 Teile verfügbar (3 folgen zeitnah)
  - Erste Forschungseinsätze
  - Erster Einsatz im Industrie-Umfeld
  - Internationales Interesse
- **Ausblick**
  - Wir suchen Projekte zur Anwendung
  - Wir suchen Gestalter für Evaluierung & Evolution







Social Card

Statement

Normung liefert die Semantik unserer Welt. Für neue Technologien ist dies die Basis für Verständnis, Sicherheit und Akzeptanz.

Dr. Henrik Putzer
Leiter TAS, fortiss GmbH
Geschäftsführer, cogitron GmbH

dke.de **#FuSiTagung21**

Bildquelle: cogitron GmbH

Your Partner
for digital Innovation.

cogitron
a mind for systems

cogitron GmbH
Stefaniweg 4
85652 Pliening
Germany

+49 15 255 90 10 40
info@cogitron.de
www.cogitron.de

HRB 234778 Munich
UStId: DE 313 565 192