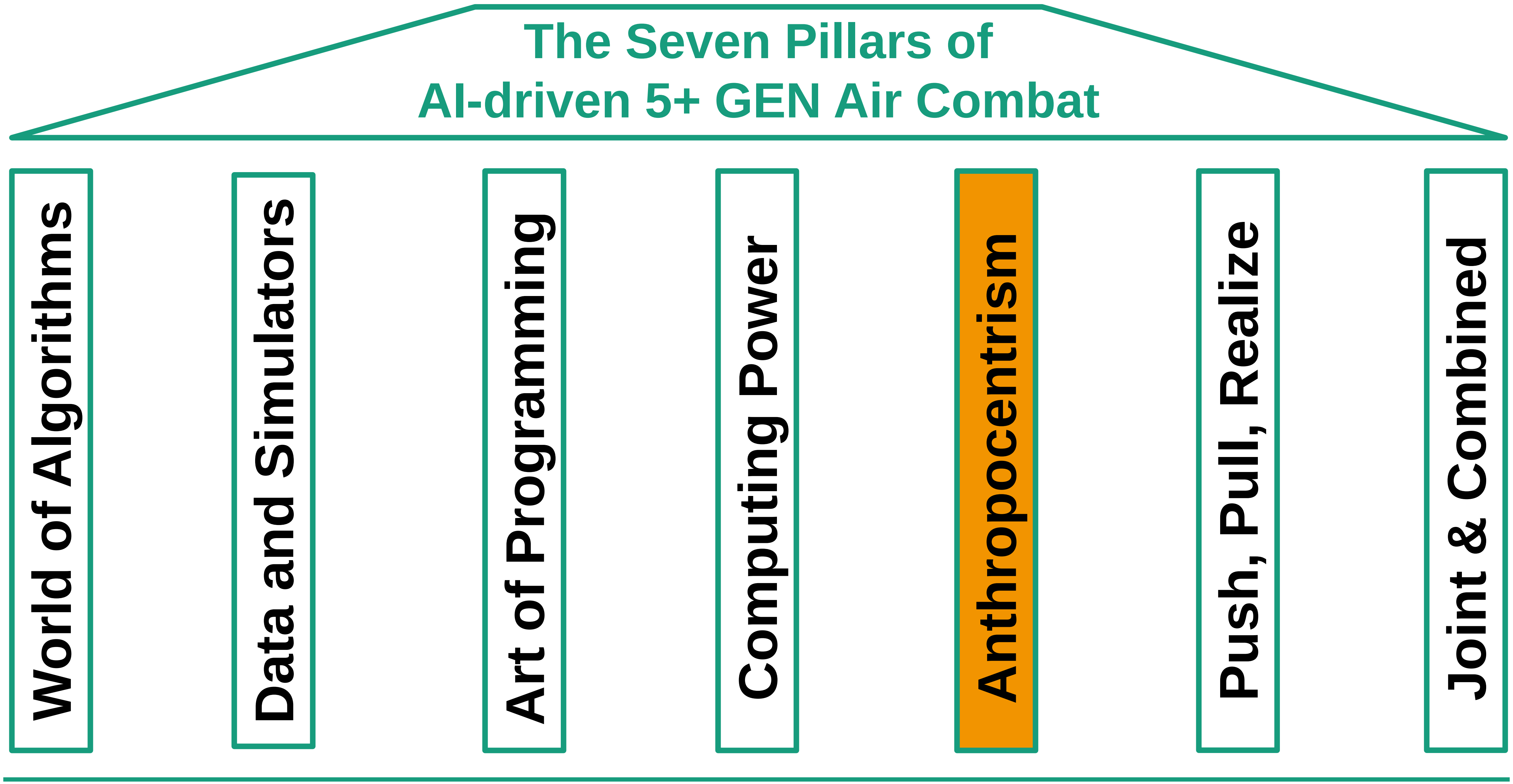
Value-Based Engineering und Standard-isierung von KI im Bereich Verteidigung

Wolfgang Koch

Prof. Dr. habil., Fellow IEEE

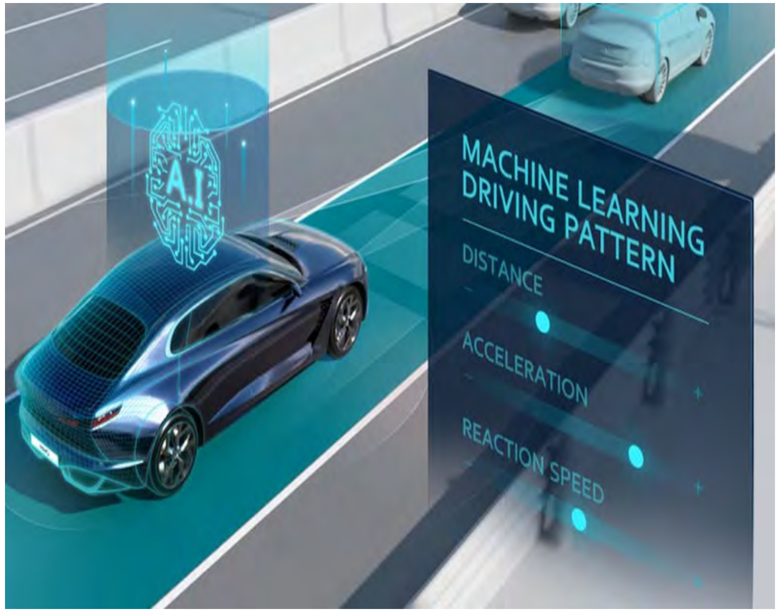
Chief Scientist, Fraunhofer FKIE



Adapted from J. E. Hyten, *Remarks To the Joint Artificial Intelligence Symposium*, Washington DC, USA, DOD, 2020

Artificially Intelligent Machines: Anthopocentric Perspective

The 'Good':
AI is assisting humans.



- Cognitive Deburdening
- Better Decision-making

The 'Bad':
Humans are assisting AI.

- Deficient Predictability
- Loss of Competency

The 'Ugly':
Humans are responsible.

Responsible humans with "de-trained" driving skills.

Künstlich-intelligente Entscheidungen zur Verteidigung

Diskussion mit General, Defense engineers und Juristen im Vorfeld des *Deutschen Juristentags* im OLG Köln am 12. 09. 2022

Bericht über die Vorfeldveranstaltung auf dem DJT

Fraunhofer
FKIE

Digitale Ethik: Nachdenken über den rechten Umgang mit digitalen Technologien. Erforderlich ist ein *Menschenbild*, das *Vernunft, Wille, Verantwortung* gedanklich ermöglicht.

Digitale Ethos betrifft die Haltung der Entscheider auf allen Ebenen; steht in engem Zusammenhang mit dem Konzept des „Staatsbürgers in Uniform".

Digitale Moral umfasst konkrete Leitlinien für den *zulass-baren* Umgang mit künstlich intelligenter Automation, im Gefecht und in der Forschung, Entwicklung, Beschaffung.

Fraunhofer FKIE

Column V: “Anthropocentrism”

verantwortlicher Entscheider

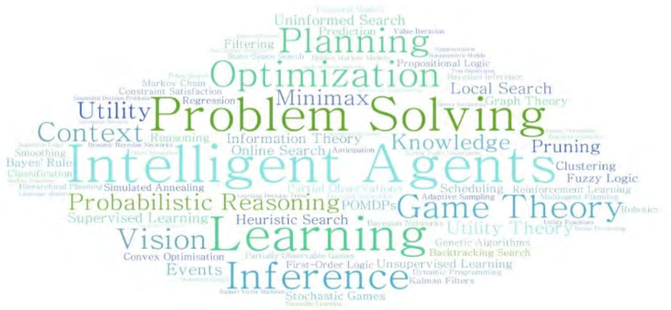
volitiv

WARUM?

kognitiv

WAS?

Technische Automation



Reife KI ist menschzentriert, modellbasiert & datengetrieben:

- prädizierbar und reproduzierbar
- Integrität: künstliche „Selbstkritik“
- insensitiv gegenüber “unknowns”
- adaptiv: variabler Einsatzkontext
- graceful performance degradation
- explainability – Basis für Zulassung
- compliance to a *code of conduct*

Künstliche Intelligenz

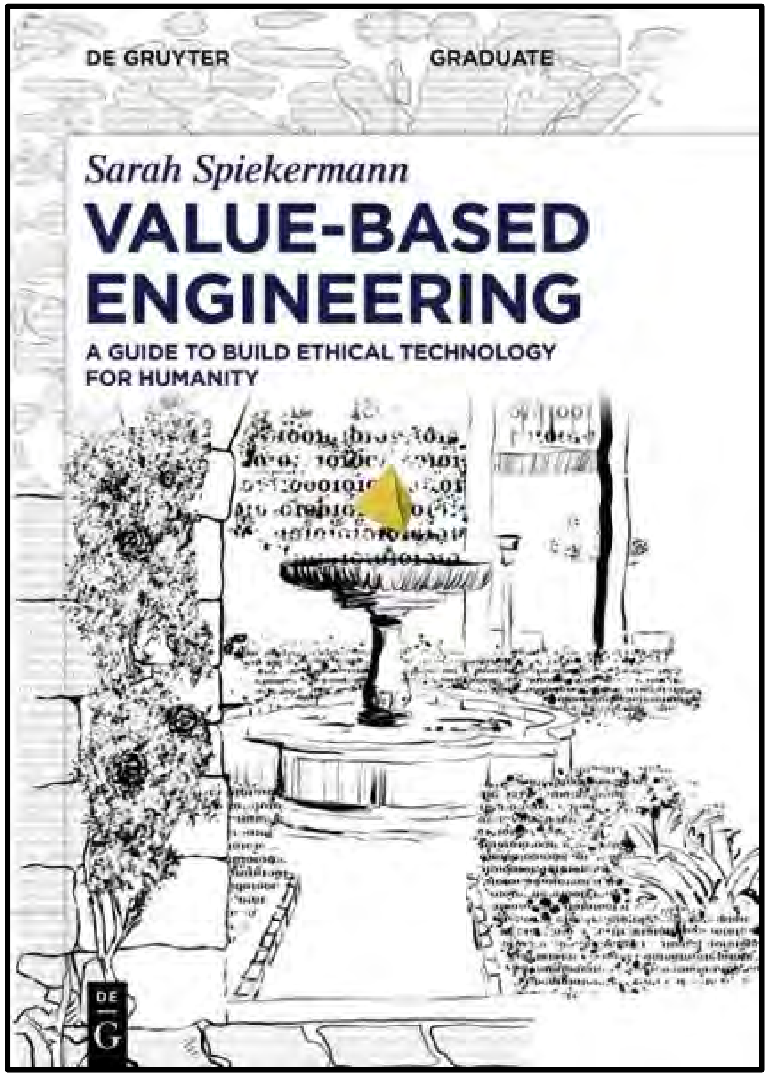

Bundesministerium der Verteidigung

Erster Bericht zur Digitalen Transformation des Geschäftsbereichs des Bundesministeriums der Verteidigung

Berlin, Oktober 2019

„Die Zukunft von KI in den Streitkräften liegt nicht in der Entscheidung Mensch oder KI, sondern in einer effektiven und skalierbaren Kombination von Mensch und KI, zur bestmöglichen Aufgabenerfüllung.“

Ziele im Operationsraum


DE GRUYTER
GRADUATE
Sarah Spiekermann
VALUE-BASED ENGINEERING
A GUIDE TO BUILD ETHICAL TECHNOLOGY FOR HUMANITY
DE G


IEEE SA
IEEE Standard Model Process for Addressing Ethical Concerns during System Design
STANDARDS
IEEE


Sarah Spiekermann
Digitale Ethik
Ein Wertesystem für das 21. Jahrhundert
DROEMER


Human nature constitutes itself (in a strong way) through its ability to *value*; that is the qualities we ”see” in the world,
how we “discern”, and whether we approach or resist.

Ideal (core) values are objective principles in the world, similar to geometric principles. They are intentional objects or abstract universals to which a species strives.

*"The ego is neither the point of departure…nor the producer of essences [like values]"**
(Max Scheler)

*Scheler, M. (1921 (1973)). Formalism in Ethics and Non-formal Ethics of Values: A New Attempt Toward the Foundation of an Ethical Personalism. USA, Northwestern University Press.

Human nature is constituted by constant valuation of the environment.

Values are phenomena that disclose what is desirable to us:

Value qualities actualize high intrinsic positive values due to value dispositions in the value bearer.

Ideal Core Value
(e.g. vitality)

Positive Value quality
(e.g. bark intactness)

Value dispositions
(Brinell x of the wood)

Value bearer:
Tree

Human nature is constituted by constant valuation of the environment.

Values are phenomena that disclose what is desirable to us:

Value qualities actualize high intrinsic positive values due to value dispositions in the value bearer.

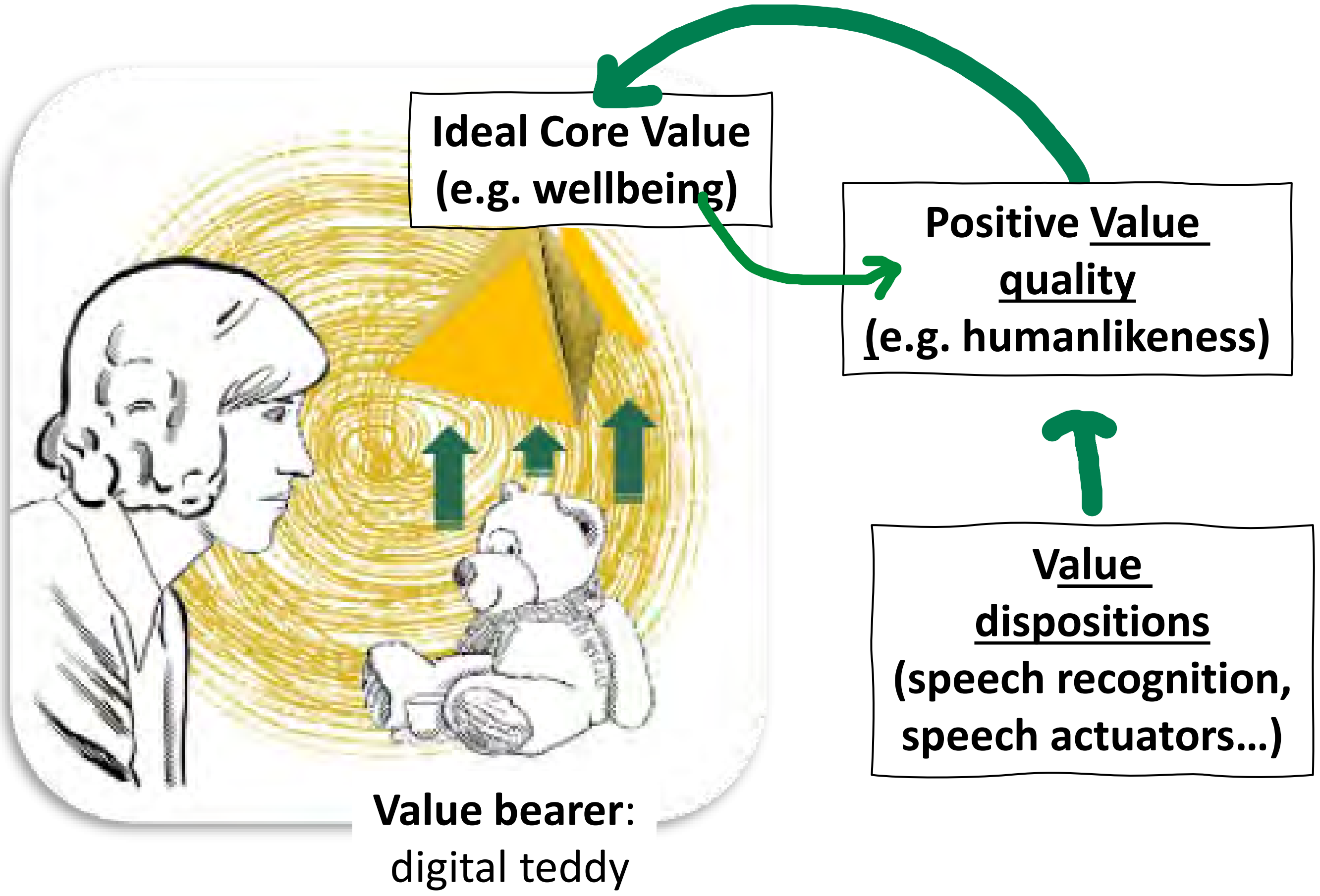
Ideal Core Value
(e.g. vitality)

Positive Value quality
(e.g. freshness)

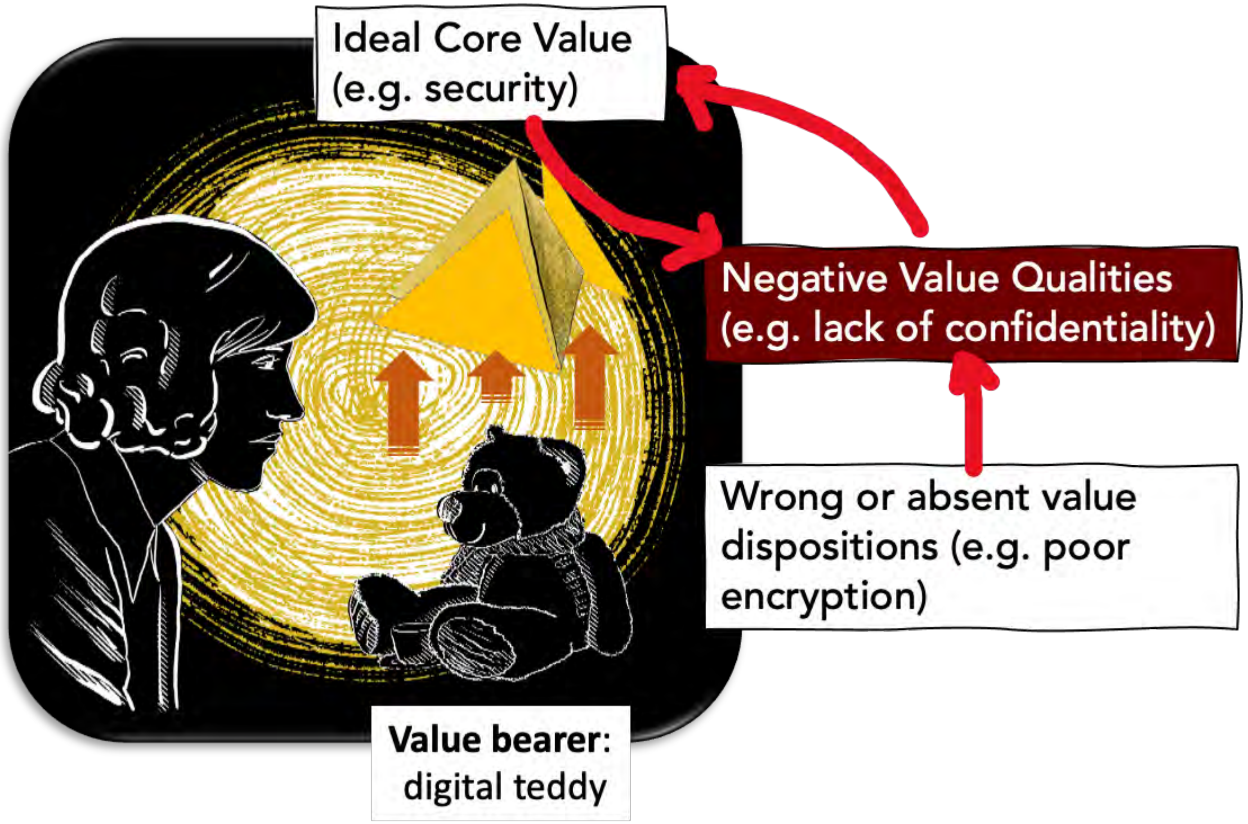
Value dispositions
(oxygen supply mechanism in tree)

Value bearer:
Forest

We can build IT systems and innovate with ideal core values in mind.



Negative value *qualities* can also be created, which undermine positive core values of a technology.



Value-based Engineering with ISO/IEC/IEEE 24748-7000

1. Ecosystem Responsibility

Value-based Engineering organizations embrace responsibility for their technical ecosystem. They abstain from partnerships or external services over which they have no control and which they cannot access.

2. Stakeholder Inclusiveness

Value-based Engineering organizations envision and plan their systems in honest and open cooperation with an extended group of direct and indirect stakeholder representatives, including critical ones.

3. Context-Sensitivity

Innovation teams in Value-based Engineering organizations deeply understand the context of their systems' deployment and anticipate its effects. In doing so they imagine what happened if one day they were a monopoly.

4. Value Identification with Moral Philosophy and/or Spiritual Tradition

Value-based Engineering organizations use moral philosophies for value elicitation, covering utilitarianism, virtue ethics and duty ethics; complemented by a culture-specific philosophical or spiritual framework from the region of the world in which a system is deployed.

5. Understanding values at depth

Value-based Engineering does not only elicit values from stakeholder concerns and context analysis, but delves deeply into them conceptually to gain a complete understanding of how they may play out in system deployment.

6. Leadership Engagement

Corporate leaders engage in introspection and support only those core values as future system principles that they would want to become universal and are therefore willing to publicly endorse.

7. Respect for Regional Laws and International Agreements

Value-based Engineering organizations respect that the ethical principles embedded in laws and signed agreements of target markets provide the outer boundary condition for their own action and therefore do not prioritize their own system values over and above these.

8. Willingness to renounce Investment

Value-based Engineering organizations actively consider not investing in a system if there are ethical grounds for such renunciation.

9. Transparency of the Value Mission

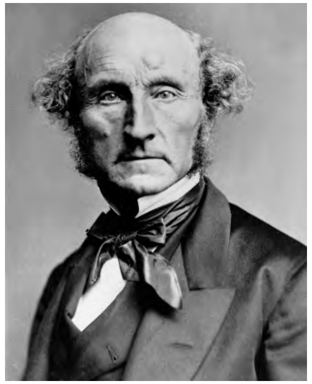
Value-based Engineering organizations publish an Ethical Policy Statement. This value-mission statement summarizes the value-priorities committed to in a system and is openly endorsed by organizational leaders. They also build up an Ethical Value Register that allows project management and auditors to recap over time what the value effects were that the system sought to cater to, and what levels of control were chosen by engineers to address likely value threats.

10. Risk-based System Design

Value-based Engineering organizations derive Ethical Value Quality Requirements (EVRs) for all core values they pursue, which then co-determine the long-term engineering roadmap. They seek to generally accommodate a “risk-thinking” in their established design and development processes.

**Kantische Maximen,
wenn rechtsbezogen**

Value Elicitation



**utilitaristische Werte,
wenn technikbezogen**

**tugendethische Werte,
wenn menschbezogen**

AI approaches that can be interpreted directly (more or less …)

Linear Models
(Logistic / linear regression)

Naïve Bayes Classifier
(based on independence assumption)

Decision Trees

k-nearest Neighbor

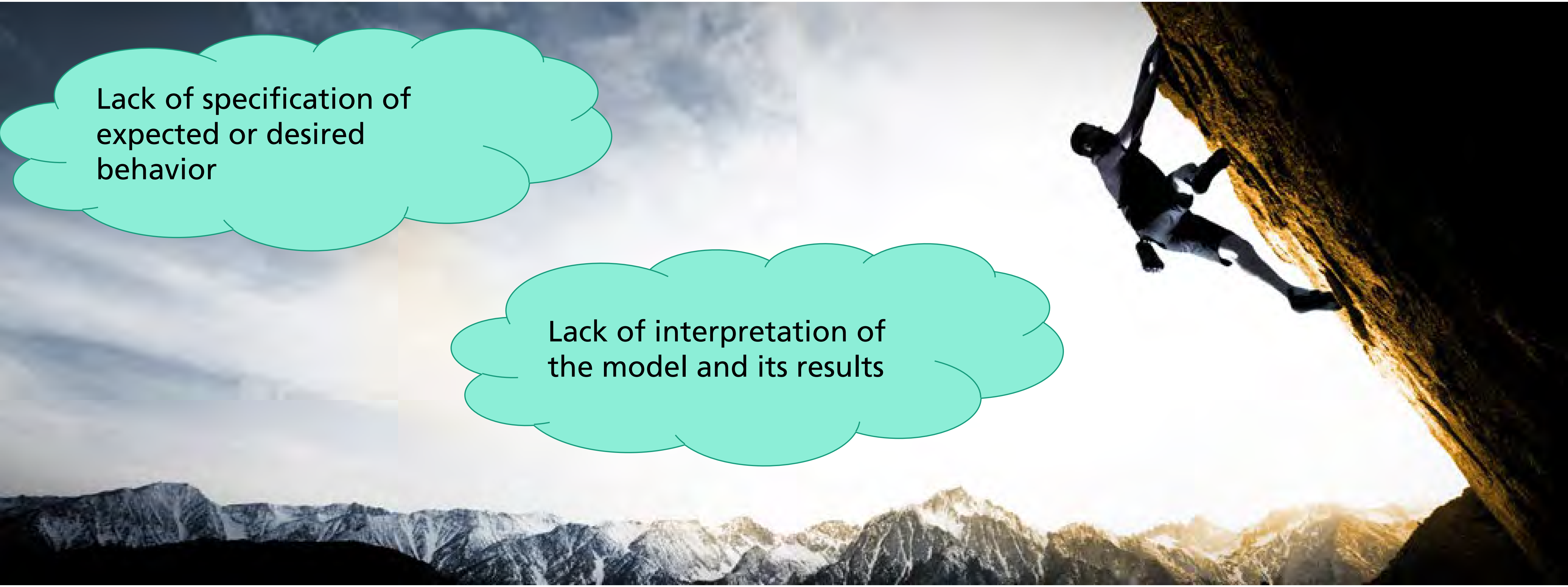
Rule based methods
(if … then … else …)

Interactive Models
(with human in the loop)

AI Certification: Schematic Procedure

- Developer demonstrates the correctness of the product to an authority.
- Developer explains the product in a transparent way: development process, test procedure
- Unforeseen behavior must be detected: identification of risks
- Definition of functional requirements to mitigate these risks
- Deriving technical requirements
- Verification (checking requirements) and validation (checking in use/experiment)

Challenges in the case of AI



Differences to Classic Software Development

Strong dependence on the underlying data

- Collecting, processing, versioning much more complex. Novel skills, methods and tools are required.
- Problems in data can lead to similar (severe) malfunctions as software errors. Software errors can be avoided in a relatively fine-grained way by unit tests**. An analog to data management is missing in practice.**

Complexity

- The components are more complex to handle than classic software components. Integration and change management are more challenging as a result
- For many methods, a mathema-tical / stochastic background is required to understand how they work.
- Classical approaches to complexity reduction (modularization, abstraction layers) are not applicable to DL.
- Extra-functional properties can only be defined and tested globally.

DL is an optimization process

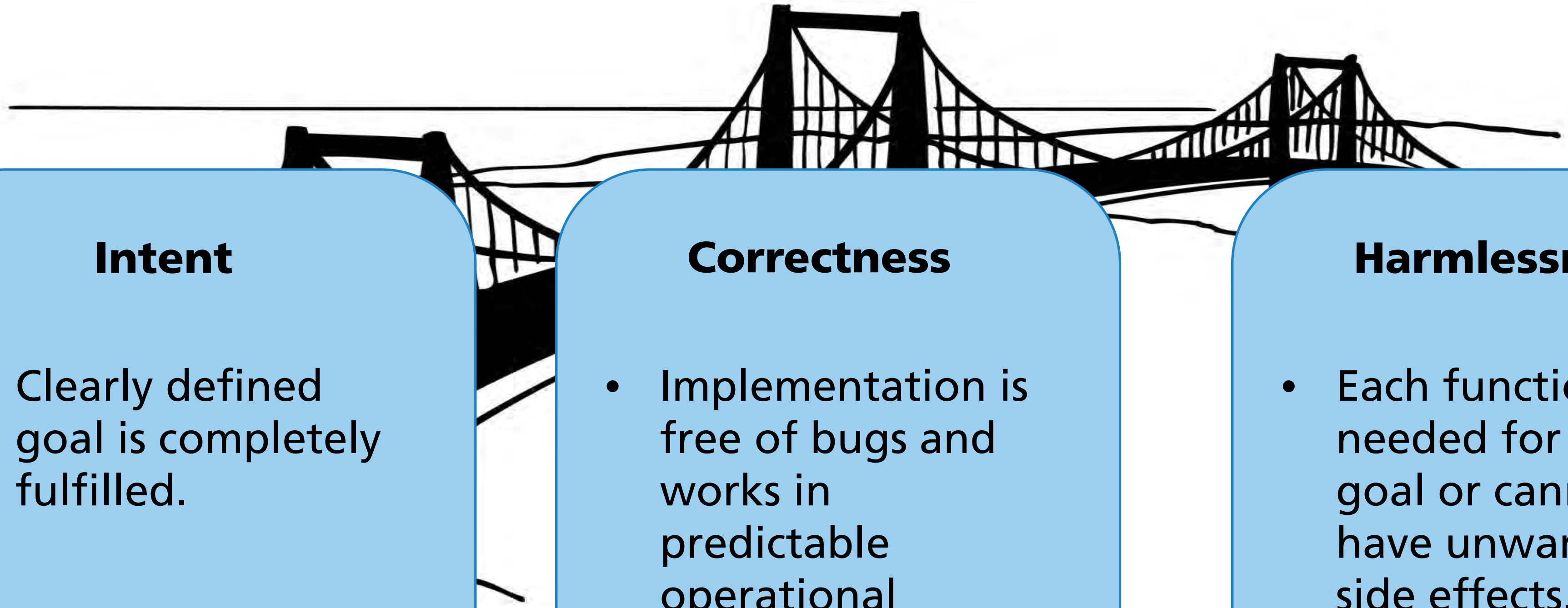
- Global optimum is unknown
- Competing goals: globalization vs. precision
- Configuration is extremely complex and often underestimated problem

intern

Fraunhofer FKIE

New approaches to certification (1): Overarching Properties (OP)

- OPs offer more freedom and less detail than verification and validation.
- However, three properties must be ensured throughout using rules:



Intent

- Clearly defined goal is completely fulfilled.

Correctness

- Implementation is free of bugs and works in predictable operational environments

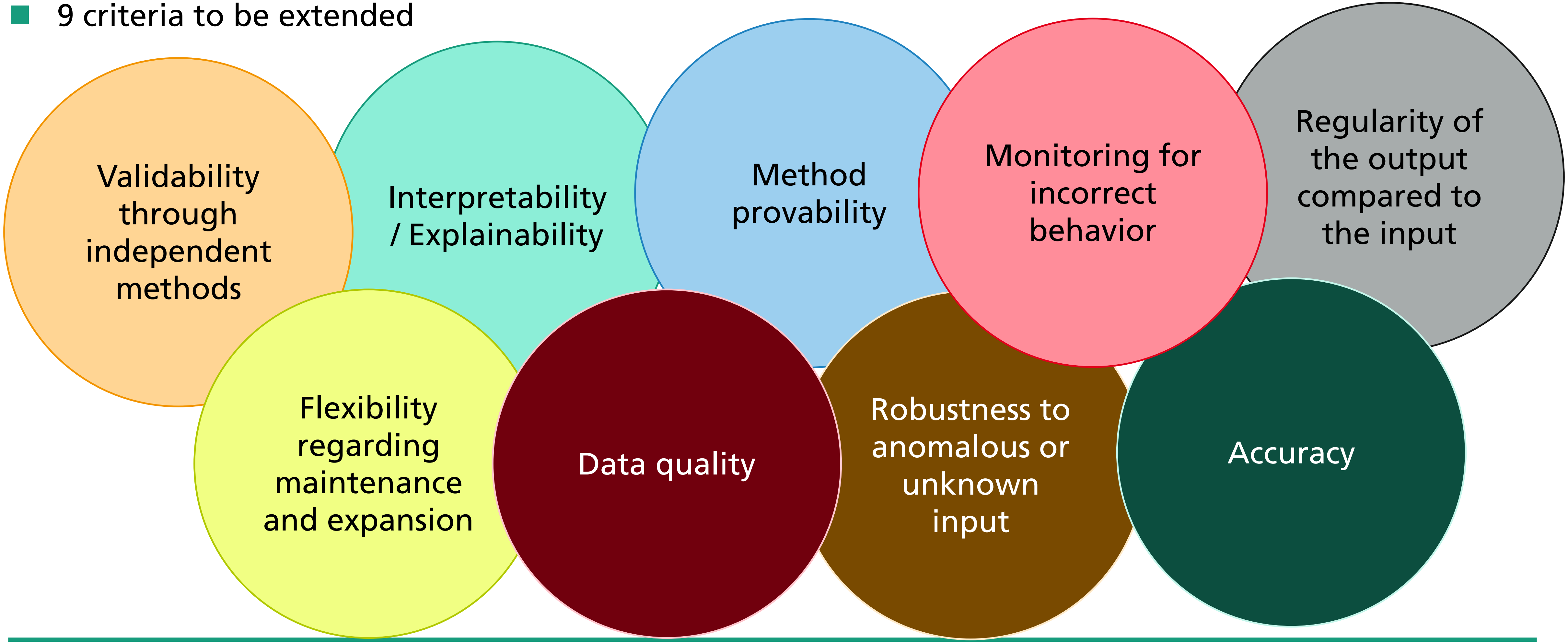
Harmlessness

- Each function is needed for the goal or cannot have unwanted side effects.

Fraunhofer FKIE

New approaches to certification (2): High Level Properties (HLP)

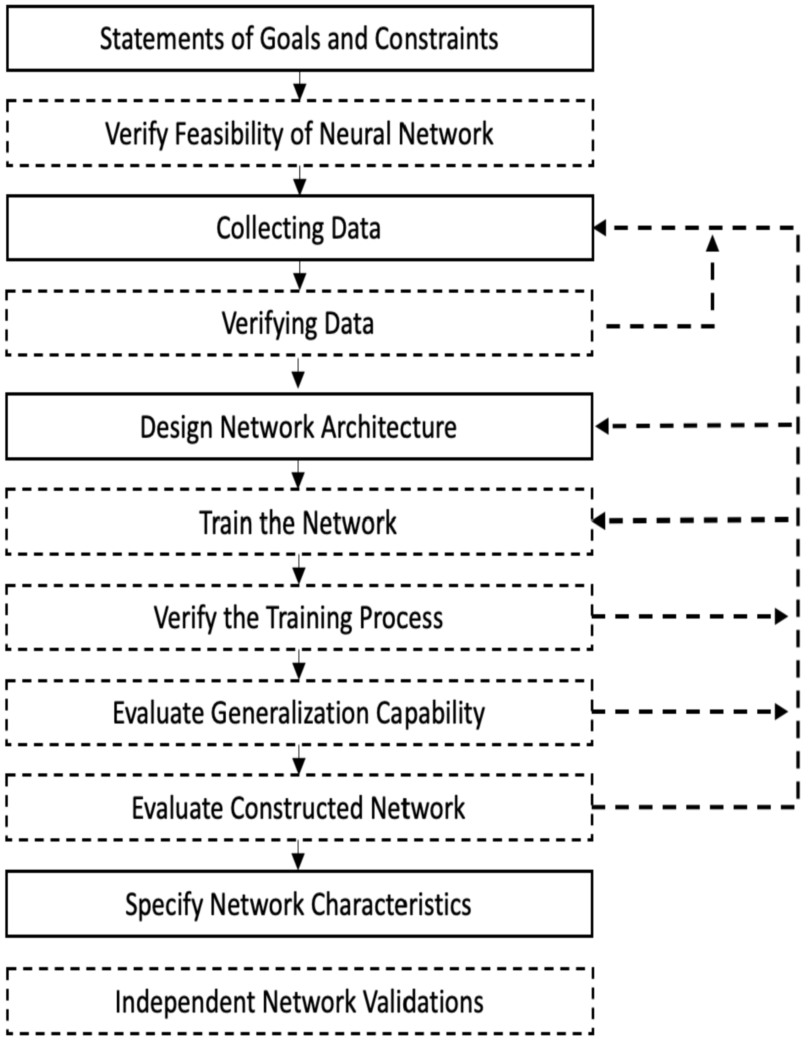
- HLPs are not related to the overall system but to the ML technique.
- 9 criteria to be extended



Security process for AI according to ISO 26262

Goes beyond ISO 26262:

- Criteria for the design and training phases must be defined.
- Ensuring robustness:
  - Sensitivity to disturbances
  - Consistency of results
- Specific process for evaluation against possible sources of error
- No end-to-end: system must be evaluable in individual functions.



© National Instruments

Fraunhofer FKIE

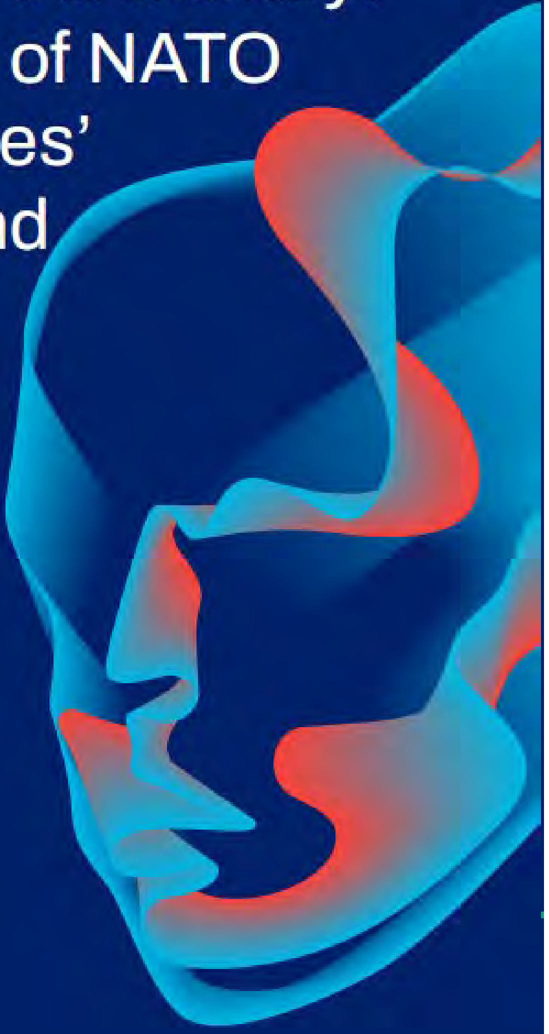
© wdrfree.com

- Certification creates transparency and trust and ensures that minimum criteria are met.
- Specifying AI behavior with any level of detail is very difficult.
- Understanding how AI works quickly becomes impossible.
- Certification processes can be applied primarily to the entire development process.
- Technical methods such as XAI and determination of variances help during development and certification.
- Ethical / legal compliance must be an integral part of certification.

CCDCOE
NATO COOPERATIVE
CYBER DEFENCE
CENTRE OF EXCELLENCE

Tallinn 2021

Artificial Intelligence and Autonomy in the Military: An Overview of NATO Member States' Strategies and Deployment



Maggie Gray, Amy Ertan
NATO CCDCOE

NATO starts work on Artificial Intelligence certification standard

07 Feb. 2023 - | Last updated: 07 Feb. 2023 19:55

English | French | Russian

NATO's Data and Artificial Intelligence Review Board (DARB) met on Tuesday (7 February 2023) to start the development of a user-friendly and responsible Artificial Intelligence (AI) certification standard to help industries and institutions across the Alliance make sure that new AI and data projects are in line with international law, as well as NATO's norms and values.