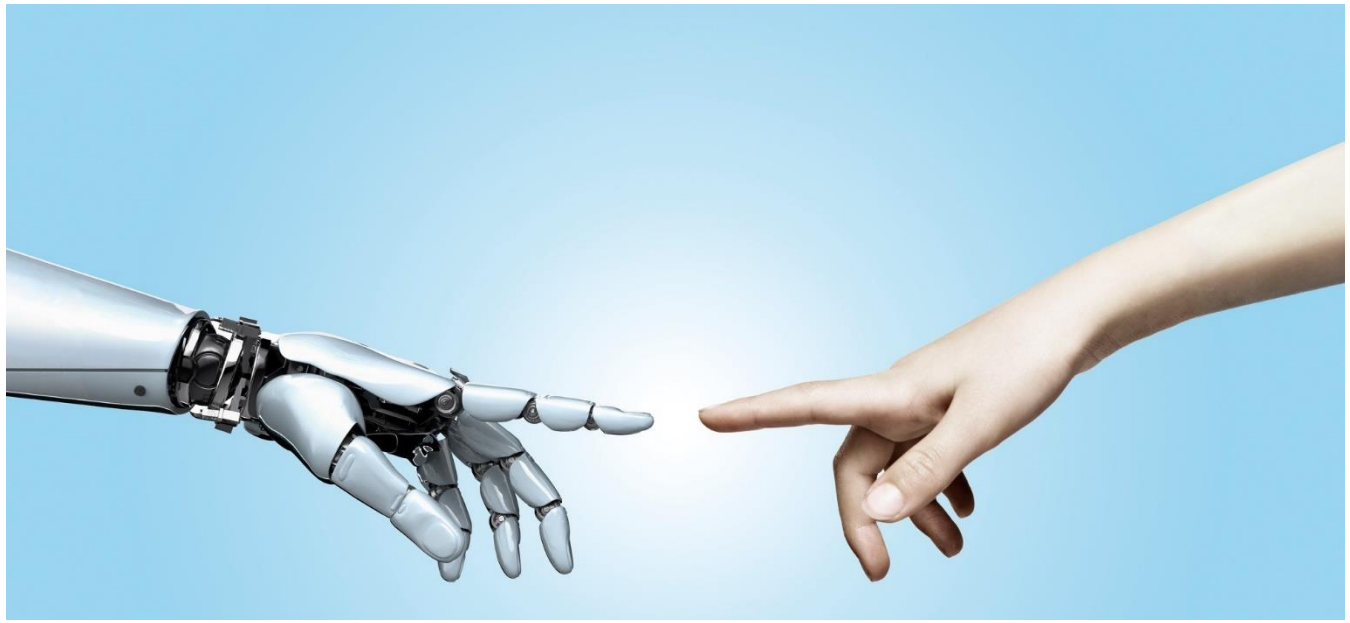
Internationale AI Standardisierung – EU Rahmen und Trustworthiness im Fokus

@Plattform Industrie 4.0
18.10.2022

Martina Paul, MBA, Vorsitz ASI AG 001.42 Artificial Intelligence
(Mirror Committee to ISO/IEC JTC 001/SC 42, -SC 22, -SC 38
CEN/CLC JTC21 Artificial Intelligence)
Rania Wazir, PhD, Mitglied ASI AG 001.42 & PE ISO/IEC 12792 „Transparency taxonomy of AI systems“

- ✓ Was verstehen wir unter vertrauenswürdiger KI
- ✓ Herausforderungen
- ✓ Der Weg der Europäischen Kommission
- ✓ Projekte
- ✓ Warum teilnehmen?



Facts & Figures

- Spiegelkomitee zu drei ISO/IEC JTC1 Subcommittees und einem CEN/CLC JTC
- 56 teilnehmende Personen
- Regelmäßige Vorträge zu Standardisierungsthemen (GAIA X, ÖCloud, AI Standardisierungsroadmap Dtl).
- Monatliche bis zweimonatliche Sitzungen

STRUKTUR

ISO/IEC JTC 1/SC 42/AG 3 ⓘ	AI standardization roadmapping
ISO/IEC JTC 1/SC 42/AHG 1	Dissemination and outreach
ISO/IEC JTC 1/SC 42/AHG 2 ⓘ	Liaison with SC 38
ISO/IEC JTC 1/SC 42/AHG 4 ⓘ	Liaison with SC 27
ISO/IEC JTC 1/SC 42/JWG 2 ⓘ	Joint Working Group ISO/IEC JTC1/SC 42 - ISO/IEC JTC1/SC 7 : Testing of AI-based systems
ISO/IEC JTC 1/SC 42/WG 1 ⓘ	Foundational standards
ISO/IEC JTC 1/SC 42/WG 2 ⓘ	Data
ISO/IEC JTC 1/SC 42/WG 3 ⓘ	Trustworthiness
ISO/IEC JTC 1/SC 42/WG 4 ⓘ	Use cases and applications
ISO/IEC JTC 1/SC 42/WG 5 ⓘ	Computational approaches and computational characteristics of AI systems

Vertrauenswürdige KI – was ist das?

Trustworthy AI – wozu?

New threats from AI:

- **Security**
- **Privacy**
- **Discrimination & bias**
- **Opaqueness**
- **Unpredictability**
- **Lack of accountability**
- **Misuse & abuse**
- **Environmental damage**

Trustworthy AI – wozu?

- Ziad Obermeyer et al. ***Dissecting racial bias in an algorithm used to manage the health of populations***. https://science.sciencemag.org/content/366/6464/447
- After Google's Gorillas comes Facebook's Primates: ***Facebook Apologizes After A.I. Puts 'Primates' Label on Video of Black Men,*** September 2021. https://www.nytimes.com/2021/09/03/technology/facebook-ai-race-primates.html
- The Guardian, ***Amazon ditched AI recruiting tool that favored men for technical jobs,*** Reuters, Oktober, 2018. https://www.theguardian.com/technology/2018/oct/10/amazon-hiring-ai-gender-bias-recruiting-engine
- Hao, K., ***Training a single AI model can emit as much carbon as five cars in their lifetimes,*** in MIT Technology Review, June 6, 2019 https://www.technologyreview.com/s/613630/training-a-single-ai-model-can-emit-as-much-carbon-as-five-cars-in-their-lifetimes/

Trustworthy AI Kriterien

ISO/IEC 22989:2022

trustworthiness

- ability to meet *stakeholder* expectations in a verifiable way

Ob ein KI System vertrauenswürdig ist oder nicht, hängt damit von den Stakeholdern ab, die mitunter auch unterschiedliche Perspektiven und Werte haben können und davon, dass es verfizierbar und validierbar ist.

Trustworthy AI Kriterien

Die EU HLEG on AI hat folgende Kriterien erstellt, basierend auf dem EU Charter of Fundamental Rights:

(1) human agency and oversight,

(2) technical robustness and safety,

(3) privacy and data governance,

(4) transparency,

(5) diversity, non-discrimination and fairness,

(6) environmental and societal well-being and

(7) accountability

Trustworthy AI Kriterien

Die *EU HLEG on AI* hat folgende Kriterien erstellt, basierend auf dem *EU Charter of Fundamental Rights*:

(1) human agency and oversight

(2) technical robustness and safety

(3) privacy and data governance

(4) transparency

(5) diversity, non-discrimination and fairness

(6) environmental and societal well-being

(7) accountability

- Was genau bedeuten diese Kriterien?
- Wie sind sie umzusetzen?
- Wie kann garantiert werden, dass ein KI System diesen Kriterien auch entspricht?

Vertrauen schaffen durch robuste Konformitätsbewertung

- ✓In ISO/IEC conformity assessment is defined as a demonstration that specified requirements relating to a product, process, system, person or body are fulfilled [ISO/IEC 17000:2020].

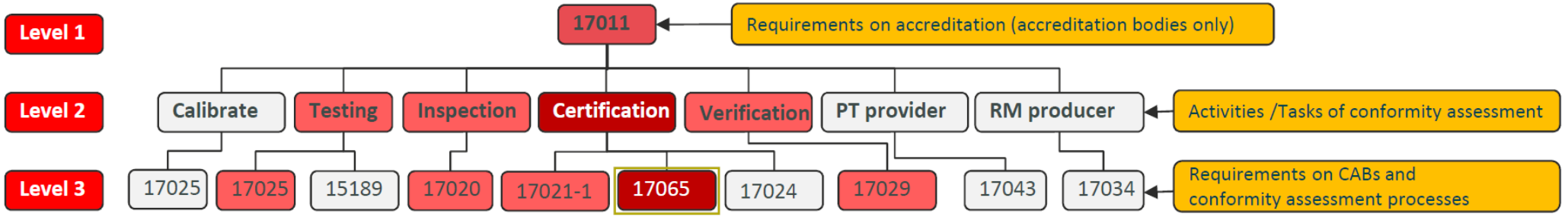
- ✓Conformity assessment procedures, such as testing, inspection and certification, offer assurance that products fulfil the requirements specified in standards and regulations:

- ✓The well described and established responsibilities within the so called Quality Infrastructure which assures that the above mentioned objectives are met need to be followed meticulously so not to create any confusion within the market

HORIZONTAL QUALITY INFRASTRUCTURE

Normative Level System

EA-MLA (EA 1/06) or IAF (IAF PR 4)



Level 4

- Conformity assessment programme/scheme (e.g. under ISO/IEC 17065,ISO/IEC 17067; ISO/IEC 17030) ← Reproducibility/ labelling/ competence etc.
- Specification of test procedures and methods (e.g. in one particular sector) ← Requirements for procedures and methods

Level 5

Requirements for the object of assessment.
Stand-alone AI-Systems and AI-Systems as part of a product or service.

(e.g. GDPR; ISO/IEC 27701 Security techniques — Extension to ISO/IEC 27001 and ISO/IEC 27002 for privacy information management — Requirements and guidelines)

ISO/IEC 17007 = **Strict separation** of requirements on the "**object of assessment**" (level 5) and "**test methods**" (levels 3 & 4).

For the activity of testing (i.e. "conformity assessment" in the form of testing/inspection/certification, etc. [Level 2 to 4]) there are international standards (ISO/IEC), which define the minimum standard for these organisations and for their testing activities. The same applies to the activities of the accreditation authorities, whose tasks and procedures are regulated in the ISO/IEC 17011 standard. The reciprocal agreements (MLA/MRA) administered by the international organisations EA, ILAC and IAF are binding under international law (international executive administrative agreement according to Art. 59 Para. 2 Sentence 2 GG), in order to achieve an international, mutual recognition of the German accreditation and conformity assessment. At the same time, this describes the minimum professional standard for such activities. Anyone who falls short of these standards does not test lege artis (= in conformity with the law).

DAkkS

What are the specificities of an AI system if applied in organizations, products, or by persons that lead us to the assumption it might require complementary determination activities than those already existing?

Criteria:

- **Approximation algorithms, output is stochastic**
- **Application in complex environments**
- **Evolving over time**
- **Application in automated decision making**

Where does a need for Conformity Assessment Schemes and respective requirement standards come from?

- **Proposal EU – Regulation (Art 40)***
- **Policy Makers in general (UK proposal)**
- **Canadian Algorithm Assessment**

* https://eur-lex.europa.eu/legalcontent/EN/TXT/?uri=CELEX%3A52021PC0206

ISO/IEC JTC1/SC 42 Standards, die fit für CA sind

- ✓ISO/IEC DIS 24029-2 Information Technology - Artificial intelligence (AI) — Assessment of the robustness of neural networks — Part 2: Methodology for the use of formal methods
- ✓ISO/IEC 24668 Information technology — Artificial intelligence — Process management framework for big data analytics
- ✓ISO/IEC DIS 5338 Information technology — Artificial intelligence — AI system life cycle processes
- ✓ISO/IEC DIS 25059 Software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Quality model for AI systems
- ✓ISO/IEC DIS 42001 Information Technology — Artificial intelligence — Management system

Beispiel

- ✓ISO 31000 – Risk management: principle on inclusiveness for risk management.
- ✓Assessment: A process is in place to ensure inclusiveness: You can check paper work, reports, names of resp. persons, etc

Examples for selection according to the **statement of conformity**:

- ✓Conformity with specified values according to a decision rule (e.g. “specification met”, “threshold passed”) would be attested by a test report (ISO/IEC 17025).
- ✓Conformity with general requirements (e.g. “safe”, “stable”) at the point in time of the examination would be attested by an inspection report (ISO/IEC 17020).
- ✓Conformity with probability and plausibility criteria at the point in time of the extrapolation would be attested by a validation statement (ISO/IEC 17029).
- ✓Conformity with safety requirements for a product line over a period of three years with annual surveillance of production and product sampling from the market would be attested by certification (ISO/IEC 17065).

Europäische/nationale Initiativen, die in CEN/CLC JTC21 einfliessen

- VCIO based description of systems for AI trustworthiness characterisation VDE SPEC 90012 V1.0 (en)

- https://www.confiance.ai/en/

- PTB Germany: **Metrology for AI in medicine**

From <https://www.ptb.de/cms/en/research-development/into-the-future-with-metrology/herausforderung-medizin/digitalization-ptbs-ai-strategy/metrology-for-ai-in-medicine-the-ptb-team-has-grown.html

Existing projects touching upon Trustworthiness (in grün)

ISO/IEC TR 24368:2022 Information technology — Artificial intelligence — Overview of ethical and societal concern

ISO/IEC TS 4213:2022 Information technology — Artificial intelligence — Assessment of machine learning classification performance

ISO/IEC TR 24029-1:2021 Artificial Intelligence (AI) — Assessment of the robustness of neural networks — Part 1: Overview

ISO/IEC TR 24027:2021 Information technology — Artificial intelligence (AI) — Bias in AI systems and AI aided decision making

ISO/IEC TR 24028:2020 Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence

DELIVERABLES UNDER DEVELOPMENT

ISO/IEC CD 5259-1 Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 1: Overview, terminology, and examples

ISO/IEC CD 5259-2 Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 2: Data quality measures

ISO/IEC CD 5259-3 Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 3: Data quality management requirements and guidelines

ISO/IEC CD 5259-4 Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 4: Data quality process framework

ISO/IEC AWI 5259-5 Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 5: Data quality governance

ISO/IEC DIS 5338 Information technology — Artificial intelligence — AI system life cycle processes

ISO/IEC CD 5339 Information Technology — Artificial Intelligence — Guidelines for AI applications

ISO/IEC CD 5392 Information technology — Artificial intelligence — Reference architecture of knowledge engineering

DELIVERABLES UNDER DEVELOPMENT

ISO/IEC CD TR 5469 Artificial intelligence — Functional safety and AI systems

ISO/IEC AWI TS 5471 Artificial intelligence — Quality evaluation guidelines for AI systems

ISO/IEC AWI TS 6254 Information technology — Artificial intelligence — Objectives and approaches for explainability of ML models and AI systems

ISO/IEC DIS 8183 Information technology — Artificial intelligence — Data life cycle framework

ISO/IEC AWI TS 8200 Information technology — Artificial intelligence — Controllability of automated artificial intelligence systems

ISO/IEC AWI TS 12791 Information technology — Artificial intelligence — Treatment of unwanted bias in classification and regression machine learning tasks

ISO/IEC AWI 12792 Information technology — Artificial intelligence — Transparency taxonomy of AI systems

ISO/IEC AWI TR 17903 Information technology — Artificial intelligence — Overview of machine learning computing devices

ISO/IEC FDIS 23894 Information technology — Artificial intelligence — Guidance on risk management

DELIVERABLES UNDER DEVELOPMENT

ISO/IEC DIS 24029-2 Artificial intelligence (AI) — Assessment of the robustness of neural networks — Part 2: Methodology for the use of formal methods

ISO/IEC AWI TR 24030 Information technology — Artificial intelligence (AI) — Use cases

ISO/IEC DIS 25059
Software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Quality model for AI systems

ISO/IEC AWI TS 29119-11
Information technology — Artificial intelligence — Testing for AI systems — Part 11:

ISO/IEC DIS 42001
Information technology — Artificial intelligence — Management system

ISO/IEC AWI 42005
Information technology — Artificial intelligence — AI system impact assessment

CEN/CLC JTC21 Artificial Intelligence - Scope

- ✓ The JTC shall produce standardization deliverables in the field of Artificial Intelligence (AI) and related use of data, as well as provide guidance to other technical committees concerned with Artificial Intelligence. The JTC shall also consider the adoption of relevant international standards and standards from other relevant organisations, like ISO/IEC JTC 1 and its subcommittees, such as SC 42 Artificial intelligence. The JTC shall produce standardization deliverables to address European market and societal needs and to underpin primarily EU legislation, policies, principles, and values

Standardization Request Überblick

- ✓Risk Management system for AI Systems
- ✓Governance and quality of datasets used to buidlAI Systems
- ✓Record keeping through logging
- ✓Transparency and information to the users of AI systems
- ✓Human oversight of AI systems
- ✓Accuracy specifications of AI Systems
- ✓Robustness specifications for AI Systems
- ✓Cybersecurity specifications for AI Systems
- ✓Quality management system for providers of AI system, including post market monitoring
- ✓Conformity Assessment for AI Systems

	WG 1: SAG Patrick Bezombes (FR) Francisco Medeiros (BE)	WG 2: Operational Aspects Emilia Tantar (LU) Ansgar Koene (UK)	WG 3: Engineering Aspects James Davenport (UK) Anders Kofod-Petersen (DK)	WG 4: Foundational & Societal Aspects Laurence Devillers (FR)
NWIPs already approved by ballot	Advisory group – does not develop standards	▪ TR **AI Conformity Assessment** (Emilia Tantar, LU)	▪ TR **Overview of AI tasks and functionalities related to natural language processing** (Lauriane Aufrant, FR)	▪ EN **AI-Enhanced Nudging** (Laurence Devillers, FR) ▪ EN **Adoption of ISO/IEC 22989 - Artificial intelligence concepts and terminology** ▪ EN **Adoption of ISO/IEC 23053 - Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)**
NWIPs in preparation	-	▪ PWI (TR) **AI Risks – Check List for AI Risks Management** (Renaud Di Francesco, IT) ▪ PWI (EN) **Adoption of ISO/IEC 42001 AI Management system** (Martha Janczarski, IE)	▪ PWI (TR) **Data Governance and data quality for AI in the European context** (Domenico Natale, IT)	▪ PWI (EN) **AI trustworthiness characterization** (Agnes Delaborde, FR / Henri Sohier, FR) ▪ TR **Green/Sustainable AI** (Valerie Livina, UK)
Expected further NWIPs from…	-	▪ SR1 (risk management system) ▪ SR2 (governance/quality of data) ▪ SR9 (quality mgmt) ▪ SR10 (conformity assessment)	▪ SR3 (logging) ▪ SR6 (accuracy) ▪ SR7 (robustness) ▪ Vertical standards ▪ Resource provisioning	▪ AHG8 work (incl. augmented goal specification) ▪ SR4 (transparency) ▪ SR5 (human oversight)
WG Secretariat	DS	BSI	DS	AFNOR

An der Standardsentwicklung für KI teilnehmen

Standardsentwicklung – warum mitmachen?

- Vielfältig
- Spannend
- Zukunft von AI mitgestalten

Wer macht bei den Standards mit?

- Universitäten & Forschungsinstitute
- Große Unternehmen & KMU
- NGOs, z.B. Konsumentenschutz Organisationen & Gewerkschaften

Wer macht bei den Standards mit?

- AI Expert*innen & ML Engineers
- Data Scientists & Statistiker*innen
- Expert*innen für Governance, Management, Auditing, Zertifizierung
- Infomatiker*innen & Software Testing
- Jurist*innen
- Ethiker*innen & Philosoph*innen
- Cognitive Scientists & Linguist*innen
- Wirtschaftswissenschaftler*innen
- Psycholog*innen & Sozialwissenschaftler*innen
- uvm

Beispiele aus der Standardsarbeit

ISO/IEC JTC1/SC 42 Artificial Intelligence

Was ist Bias?

Oxford English Dictionary 1:

Bias

- *Inclination or prejudice for or against one person or group, especially in a way considered to be unfair.*

Oxford English Dictionary 2:

Bias

- *A systematic distortion of a statistical result due to a factor not allowed for in its derivation.*

Was ist Bias?

..... 1,5 Jahre + ...

Was ist Bias?

ISO/IEC TR 24027:2021 *"Bias in AI systems and AI aided decision making"*

Bias
systematic difference in treatment of certain objects, people or groups in comparison to others

- Note 1 to entry: Treatment is any kind of action, including perception, observation, representation, prediction or decision.

Definition adopted in ISO/IEC 22989:2022

"Artificial intelligence concepts and terminology"

Was ist der Zusammenhang zwischen „Quality“ & „Trustworthiness“ von AI Systemen?

ISO/IEC 22989:2022

trustworthiness

- ability to meet *stakeholder* expectations in a verifiable way

ISO 13628-2:2006

quality

- conformance to specified requirements

Transparency

- Was ist Transparency von einem AI System?
- Betrachte ich das System als ganzes, oder nur die Komponente? Welche Komponente?
- Ist Transparenz einfach nur Dokumentation, oder beinhaltet es auch Zugriff auf Code, oder auf bestimmte APIs?
- Was muss in der Dokumentation enthalten sein, damit ein AI System als “transparent” gilt?
- Wie verändern sich die Anforderungen an Transparenz mit dem Lebenszyklus eines AI Systems?
- Wieviel Transparenz, von wem, gegenüber wem?
- Kann zu viel Transparenz schädlich sein? Kommt es, z.B. mit anderen Werten wie Security und Privacy in Konflikt?
- Wie hängt Transparenz mit anderen Konzepten – wie Explainability, Reproducibility, oder Accountability zusammen?

Beispiele aus der Standardisierungssarbeit

CEN/CENELEC JTC21 Artificial Intelligence

Accuracy and AI

Im draft AI Act kommt das Wort “Accuracy” 13 Mal vor (7 im Preamble, 6 in den Artikeln).

Aber … was genau bedeutet Accuracy?

Im draft AI Act, wird es nicht definiert.

Accuracy and AI

Im draft AI Act kommt das Wort “Accuracy” 13 Mal vor. Aber … was genau bedeutet accuracy?
Im Act, wird es nicht definiert.

Warum ist eine gute Definition wichtig?

ISO/IEC TS 4213:2022 Information technology — Artificial intelligence — Assessment of machine learning classification performance

accuracy

- number of correctly classified samples divided by all classified samples
- Note 1 to entry: It is calculated as $a = (T_P + T_N) / (T_P + F_P + T_N + F_N)$

Aber: diese Definition betrifft nur Klassifikations-systeme.

- **was passiert mit, z.B. einer Regression? Oder bei unsupervised learning?**
- **auch wenn es sich um ein Klassifikations-system handelt: ist Accuracy die richtige Metrik, um Qualität der Risultate zu messen? (Man stelle sich ein Diagnostiksystem für eine seltene Krankheit vor, die in nur 0,01% der Bevölkerung auftritt. Dann wäre ein System, das bei allen Menschen „negativ“ vorhersagt zu 99,99% „accurate“. Ist es aber ein gutes System?)**

Accuracy and AI: Warum ist eine gute Definition wichtig?

ISO/IEC TS 4213:2022

accuracy

- number of correctly classified samples divided by all classified samples
- Note 1 to entry: It is calculated as $a = (T_P + T_N) / (T_P + F_P + T_N + F_N)$

ISO/IEC TS 5723:2022 Trustworthiness — Vocabulary

accuracy

- measure of closeness of results of observations, computations, or estimates to the true values or the values accepted as being true

Diese zweite Definition ist etwas breiter gefasst – macht es aber nicht unbedingt klarer, wie zu messen ist.

Ausserdem: was wäre, z.B., die Accuracy von einem Machine Translation System??

Entwurf einer EU KI Regulierung - BEISPIEL

Article 10: Data and Data Governance

Datensätze von hoch-risiko KI Anwendungen sollen folgende Anforderungen erfüllen:

- Qualitativ hoch-wertig (relevant, repräsentativ, komplett, fehlerfrei)
- Training und Testdaten unterliegen einem Data Governance und Management Regime, welches den Datenverwertungszyklus, angefangen mit der Datensammlung, über Pre-processing, Labelling, bis hin zum Einsatz in der Modellierung und Modellbewertung, nachvollziehbar und transparent macht.
- In der vorgesehenen technischen Dokumentation sollen auch Datenblätter beinhaltet sein, die die eingesetzten Daten ausführlich beschreiben. (Annex IV)

Entwurf einer EU KI Regulierung - Beispiel

Article 10: Data and Data Governance

Datensätze von hoch-risiko KI Anwendungen sollen folgende Anforderungen erfüllen:

Fragen, die mit Standards beantwortet werden könnten:

- Qualitativ hoch-wertig: was bedeutet das?
- Was sind, z.B. repräsentative Daten?
- Wie schaut ein Datenblatt aus, und was muß es beinhalten?

FRAGEN?

Veranstaltungshinweis
ASI Webinar

https://www.austrian-standards.at/de/shop/ai-act-und-standardisierung~p26638

INFO-WEBINAR

AI-ACT UND STANDARDISIERUNG

VORAUSSETZUNG FÜR DEN EINSATZ VON VERTRAUENSWÜRDIGER KI

Seminarnummer 2201060

Termin: 10.11.2022, 16:00 - 17:30 Uhr

Ort: nur virtuell

DANKE FÜR IHRE AUFMERKSAMKEIT!

Martina Paul, MBA
www.linkedin.com/in/martinapaul1

Rania Wazir, Ph.D.
www.linkedin.com/in/raniawazir