Zur Normungsroadmap KI Ausgabe 2 – 2023-07-07, Start um 13:30 Uhr

KI – bestens geschützt mit Normen und Standards

Workshop zu Normungs- und Standardisierungsbedarfen aus der Roadmap

Mit Annegrit Seyerlein-Klug und Dr. rer. nat. Henrich C. Pöhls
Moderiert von Jan Rösler

Agenda

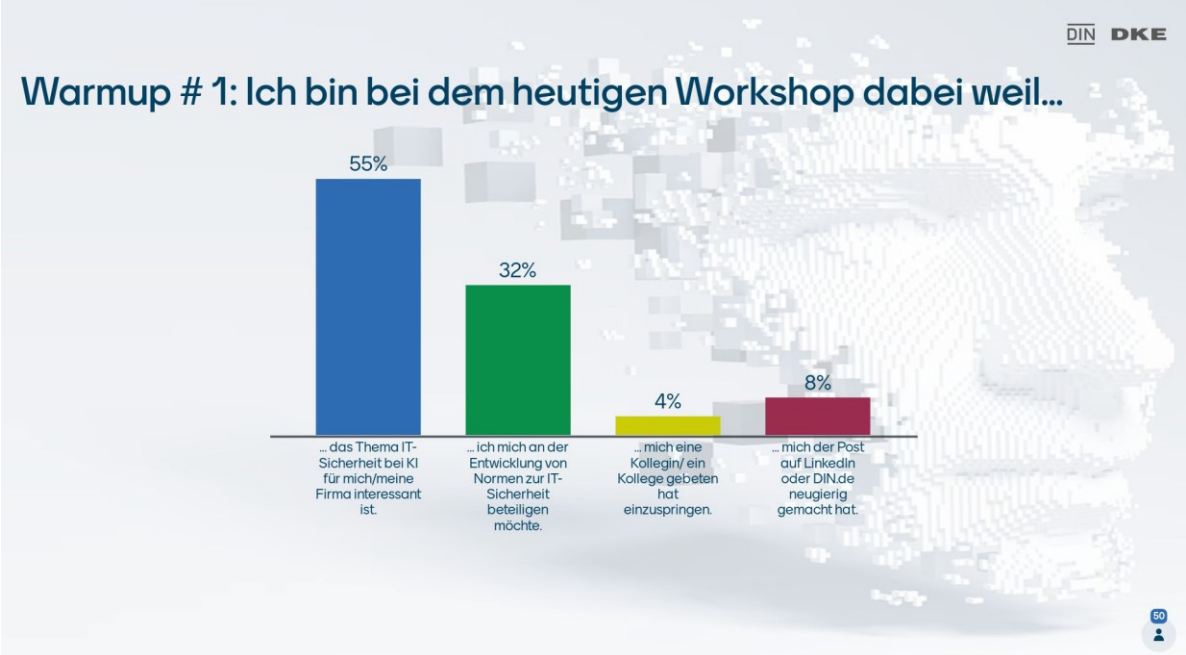
1. **Begrüßung**
   - Warmup
   - Ziele dieses Workshops
2. **Thematische Einführung**
   - Überblick zur Normungsroadmap KI
   - Vorstellung der Speaker
   - Impulsvorträge
   - Vorstellung der Normungs- und Standardisierungsbedarfe
   - Beantwortung von Fragen zu den Bedarfen
3. **Interaktive Priorisierung der Bedarfe**
   - Interessierte für Bedarfsumsetzung
   - Priorisierung der Bedarfe für den Workshop

KURZE PAUSE

4. **Diskussion der Bedarfe anhand der Priorisierung**
   - Konkretisierung der Bedarfsschwerpunkte
   - Bedarfsabgrenzung durch Anwendungsbereiche (Scopes)
   - Beziehung zwischen Bedarfen und laufenden Projekten oder veröffentlichten Normen und Standards
5. **Nächste Schritte**

1. Begrüßung

Warmup per Mentimeter-Umfrage – Ergebnisse



Mehrfachauswahl war möglich

Statements wurden von 1 bis 10 gewichtet

1. Begrüßung

Ziele dieses Workshops

Recap

- Worum geht’s überhaupt?
- Ergebnisse der Normungsroadmap KI
- Umsetzung der Handlungsbedarfe

Interaktion

- Vorstellung der Standardisierungsbedarfe
- Gemeinsames Verständnis schaffen
- Diskussion der Bedarfe

Beginn…

- der Bedarfsumsetzung und
- Vorbereitung für die Normung

Veröffentlichung der Normungsroadmap KI Ausgabe 2

- Maßnahme der KI Strategie der Bundesregierung
- Veröffentlicht am 9. Dezember 2022
- Fortschreibung der ersten Ausgabe (2020)
- Fokus: AI Act
- Kostenfreier Download:

www.din.de/go/normungsroadmapki

Filiz Elmas, Leiterin Strategische Entwicklung Künstliche Intelligenz bei DIN, Christoph Winterhalter, Vorsitzender des Vorstandes von DIN, Robert Habeck, Vizekanzler und Bundesminister für Wirtschaft und Klimaschutz, Prof. Dr. Wolfgang Wahlster, CEA des DFKI und Michael Teigeler, Geschäftsführer DKE (v.l.n.r.) © Stefan Zeitz

Überblick zur Normungsroadmap KI Ausgabe 2

Normen und Standards ermöglichen eine **zuverlässige und sichere Anwendung** von KI-Technologien und tragen zur **Erklärbarkeit und Nachvollziehbarkeit** bei.

Sie sind dadurch Schlüsselfaktoren für die **Akzeptanz von KI-Anwendungen** und schaffen **Vertrauen** am Markt und bei Verbraucher*innen.

Die Normungsroadmap KI beschreibt das **Umfeld der KI-Standardisierung** und gibt einen Überblick über bestehende Normen und Standards.

Der Fokus ist das Aufzeigen von **Normungs- und Standardisierungsbedarfen**, sowie konkreter **Handlungsempfehlung**.

Die Erarbeitung erfolgte unter Mitwirkung von Expert*innen aller relevanten Kreise.

Dafür wurden Arbeitsgruppen zu verschiedenen Schwerpunktthemen gegründet.

Bearbeitungszeitraum war von Januar 2022 bis September 2022.

Überblick zur Normungsroadmap KI Ausgabe 2

570 Fachleute

9 Arbeitsgruppen/ Schwerpunkte

18 Mitglieder Koordinierungsgruppe

318 Autor*innen

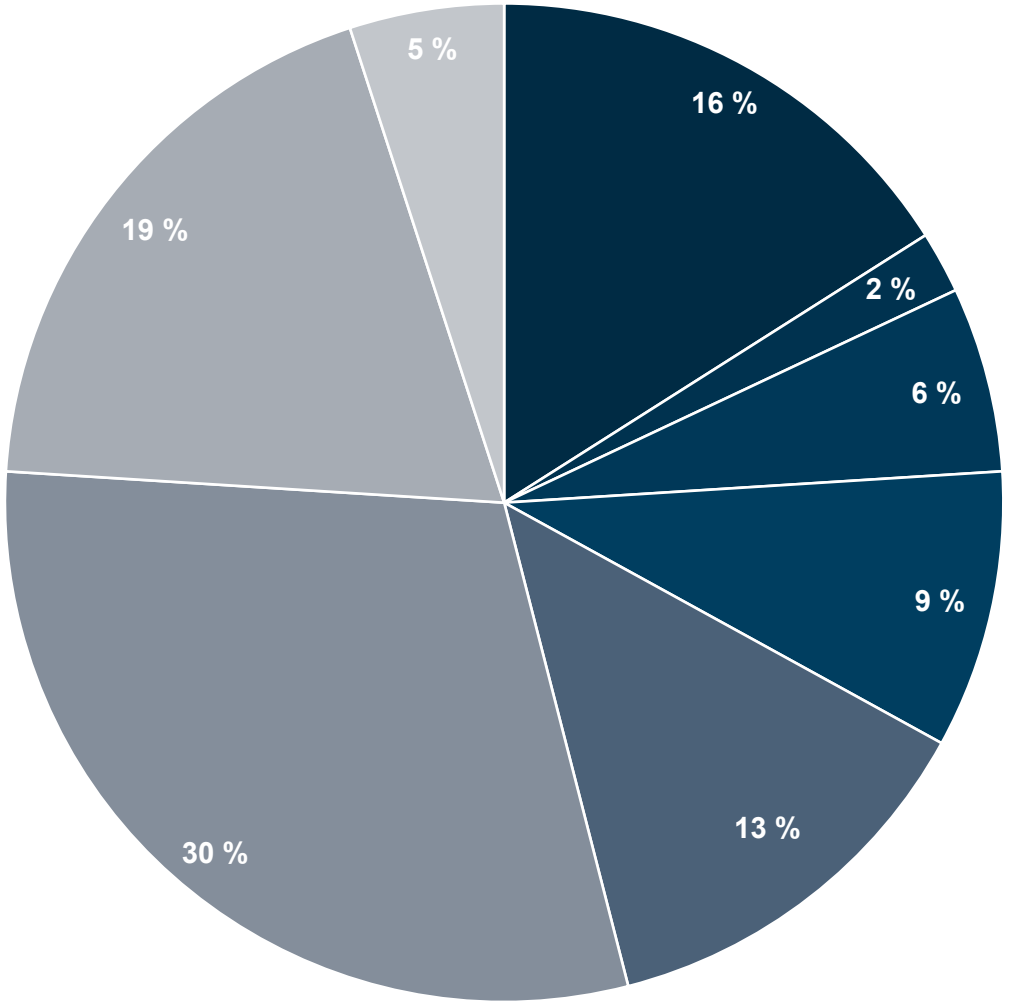
116 Handlungsbedarfe

6 Übergreifende Empfehlungen

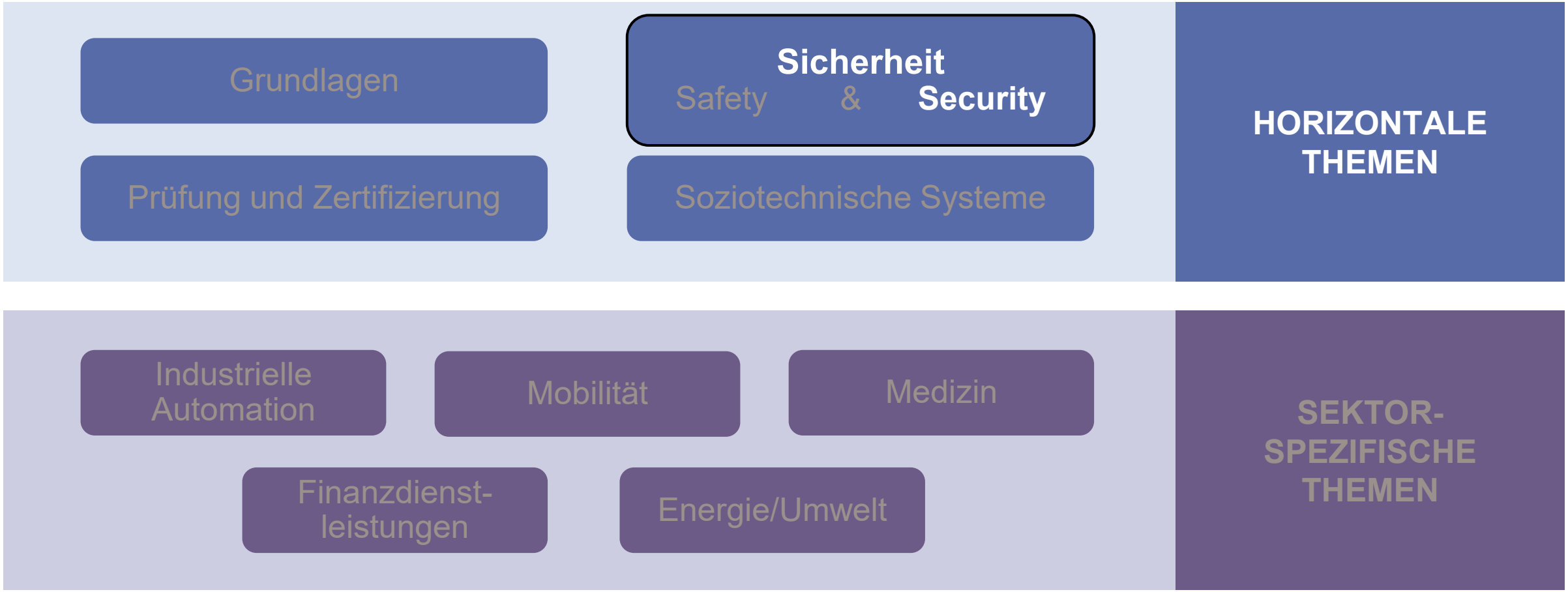
448 Seiten

Überblick zur Normungsroadmap KI Ausgabe 2

Anzahl: 570 Fachleute



Überblick zur Normungsroadmap KI Ausgabe 2



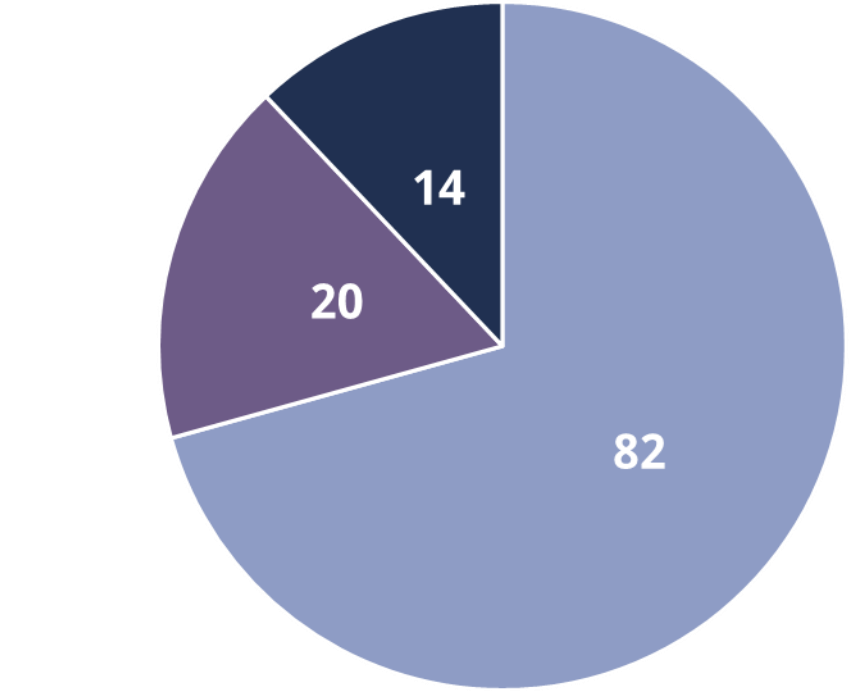
Überblick zur Normungsroadmap KI Ausgabe 2

Insgesamt 116 Handlungsbedarfe identifiziert, unterteilt in 3 Kategorien:

KATEGORIE 1:
Bedarf adressiert Normung und Standardisierung

KATEGORIE 2:
Bedarf adressiert Forschung

KATEGORIE 3:
Bedarf adressiert Politik/Gesetzgeber



Stand: Mai 2023

Conceptboard

Während des Workshops wurde Conceptboad verwendet. Das Board ist nun geschlossen. Sie können Sich aber noch per E-Mail an Jan.Roesler@din.de für die Umsetzung der Bedarfe melden – bitte Bedarfs-Code mit angeben.

Vorstellung der Speaker

Annegrit Seyerlein-Klug

Technische Hochschule Brandenburg
neurocat GmbH

Dr. rer. nat. Henrich C. Pöhls

Universität Passau

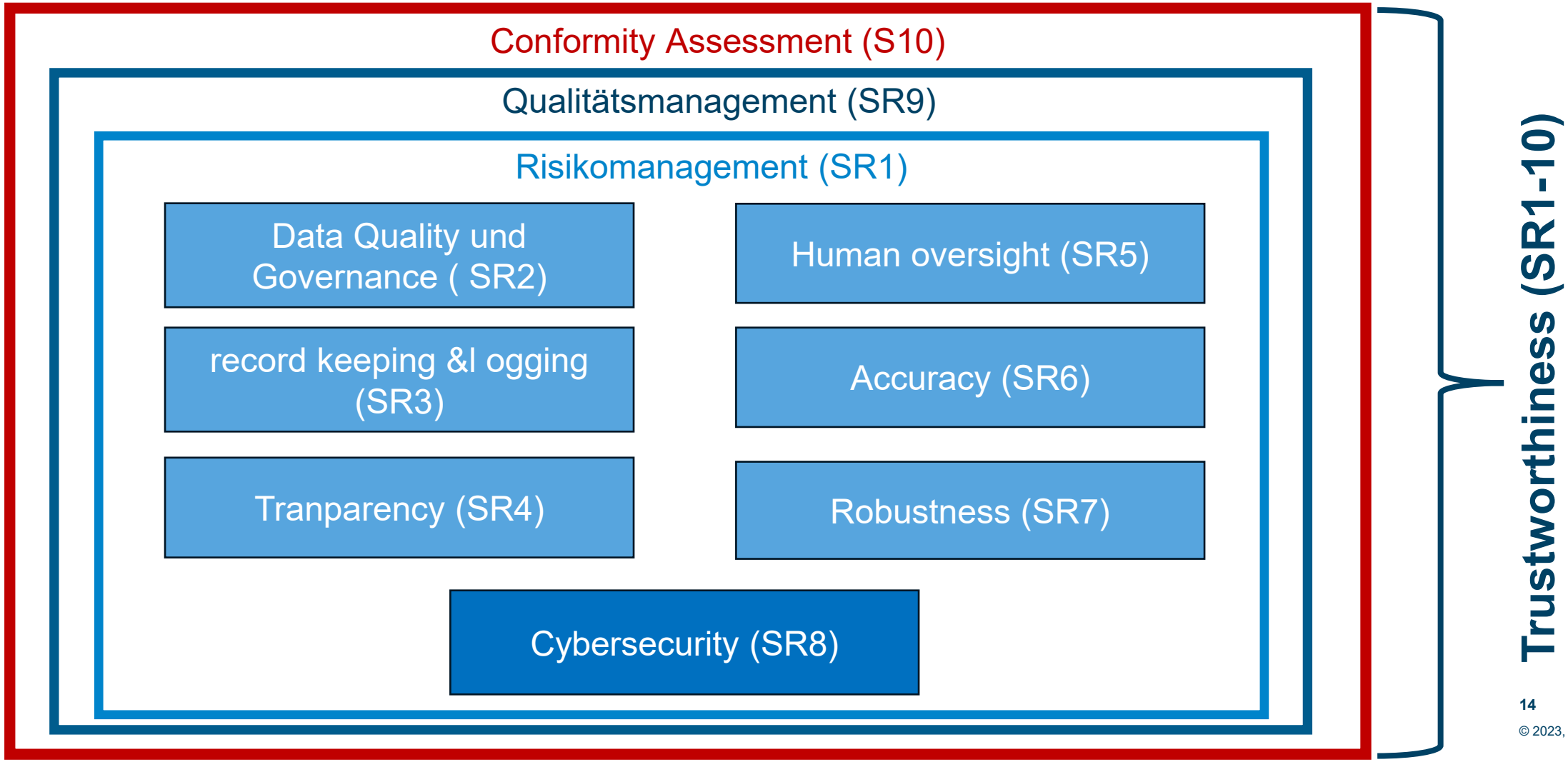
Vertrauen in KI durch Prüfung/Zertifizierung

Informations-Sicherheit bei KI und zugehörige Schutzziele, 3 Perspektiven:

- Security - Vertraulichkeit, Integrität, Verfügbarkeit, Authentizität
- Safety - Leib und Leben
- Privacy - personenbezogener Daten

Konformität
mit dem geplanten EU AI Act (Trilog) bei Hochrisikosysteme

Standardisierungsanforderung des EU AI Act



Quelle: Standardisierungsrequest AI Act

Cybersecurity specifications AI Systems (Software)

„*This European standard or European standardisation deliverable **shall provide suitable organisational and technical solutions**, to ensure that AI systems are resilient against attempts to alter their use, behaviour, or performance or to compromise their security properties by malicious third parties exploiting the AI systems' vulnerabilities.*

Organisational and technical solutions shall therefore include, where appropriate,
***measures to prevent and control cyberattacks trying to manipulate AI specific assets**, such as*

- ***training data sets** (e.g. data poisoning) or*
- ***trained models** (e.g. adversarial examples), or*
- ***trying to exploit vulnerabilities** in an AI system's digital assets or the*
- ***underlying ICT infrastructure**.*

These technical solutions shall be appropriate to the relevant circumstances and risks.

This European standard or European standardisation deliverable shall take due account of the
essential requirements for products with digital elements
*as listed in Sections 1 and 2 of Annex I to the proposal for a Regulation of the European Parliament and the Council on **horizontal cybersecurity requirements** for products with digital elements.*" – Standardization Request zum AI-Act

Quelle: Standardisierungsrequest AI Act

AI Act – Conformity Assessment for AI systems

„This European standard or European standardisation deliverable shall provide ***procedures and processes*** *for conformity assessment activities related to AI systems and* ***quality management systems*** *of AI providers.*

This European standard or European standardisation deliverable shall also provide ***criteria for assessing the competence of persons*** *tasked with the conformity assessment activities.“*

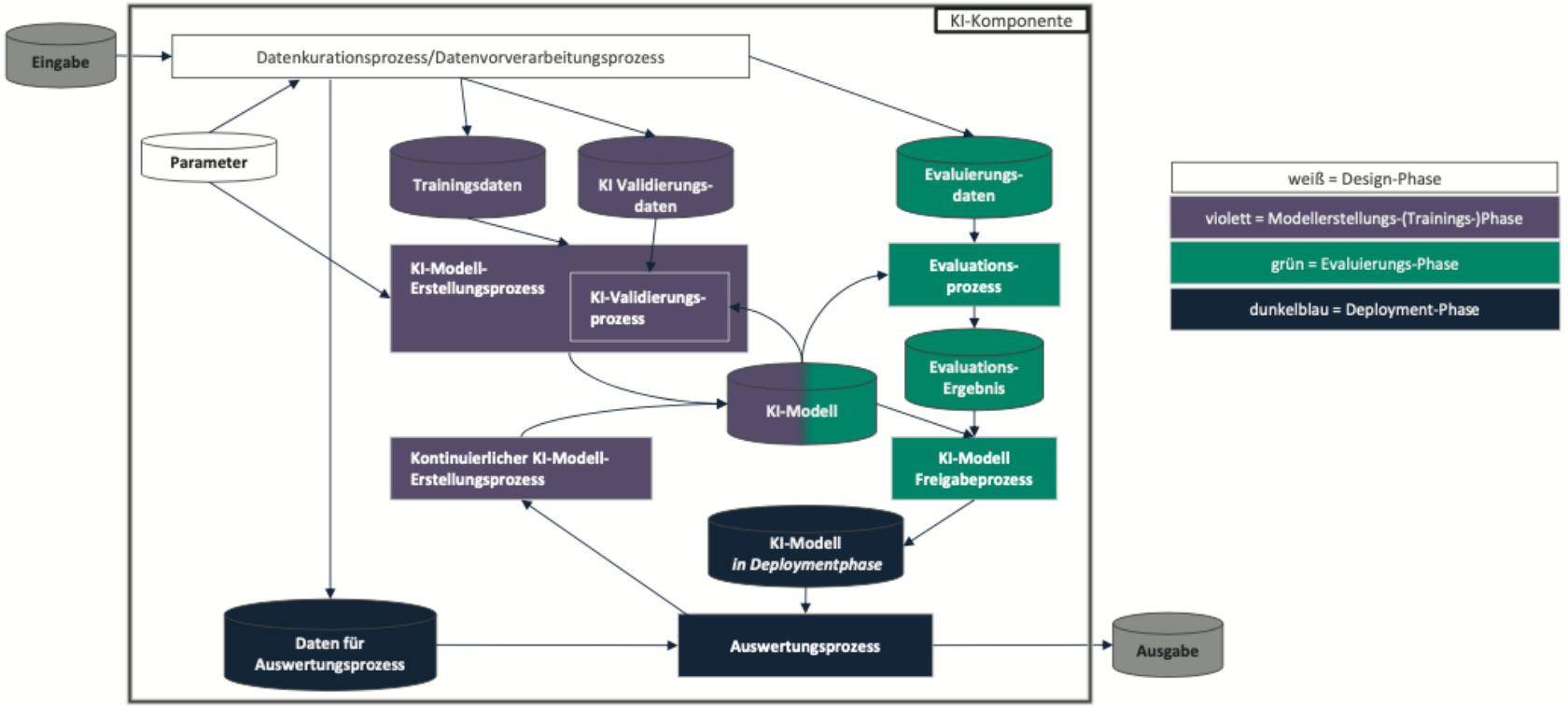
„This European standard or European standardisation deliverable shall consider both the scenarios whereby the conformity assessment is carried out by the ***provider itself*** *or with the* ***involvement of a professional external third-party organisation.***“

Security für KI – neue Herausforderungen

- Grundsätzlich können bei KI-Systemen, wie bei anderen IT-Systemen auch, alle üblichen Sicherheitsschutzziele wie Vertraulichkeit, Verfügbarkeit, Integrität oder auch Belastbarkeit durch Angriffe gefährdet sein.
- Die ‚Unzulänglichkeiten', welche spätere Angriffe ermöglichen, sind vielfältig
  - z.B. Fehler in den Daten durch menschliches Fehlverhalten
  - z.B. gezielte Manipulation von Daten (z.B. Data Poisoning von Trainingsdaten)
  - z.B. unzureichender Trainingsprozess beim Maschinellen Lernen
- Grundsätzlich drei Fragestellungen:
  1. Welche bereits existierenden Normen kann man zur Erhöhung (und zum Nachweis dieser Erhöhung) der IT-Sicherheit von KI-Systemen sinnvoll direkt nutzen?
  2. Welche bereits existierenden Normen muss man um KI-spezifische IT-Sicherheitsprobleme erweitern?
  3. Welche neuen Normen muss man für KI-spezifische Sicherheitsprobleme schaffen?

Security für KI – neue Herausforderungen

1. Herausforderung: Definition von Schutzzielen auf der Ebene von Prozessen und Daten innerhalb der KI-Komponente



NRM KI v2 – Seite 122 – Abbildung 25

Ein Beispiel: Sicherung der Trainingsdaten

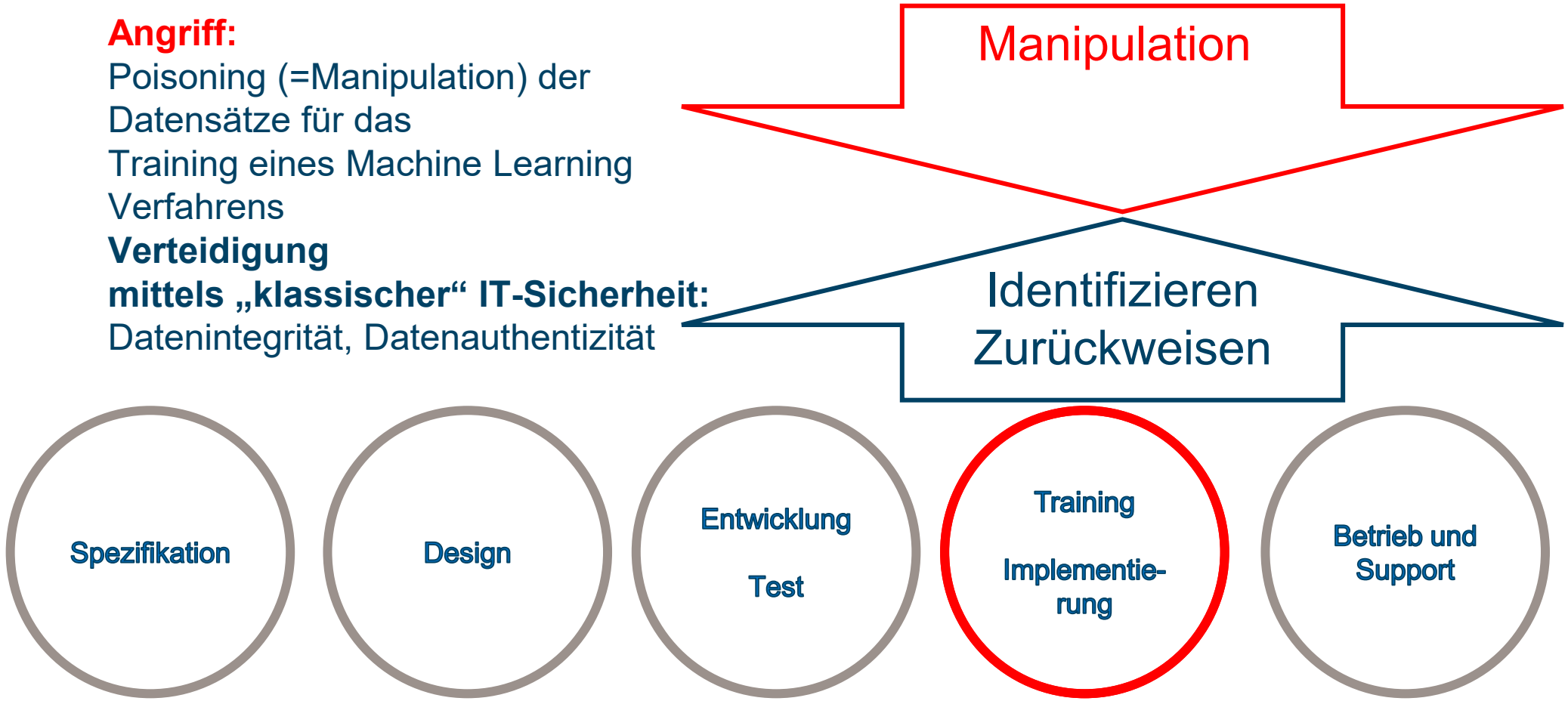
Angriff:
Poisoning (=Manipulation) der Datensätze für das Training eines Machine Learning Verfahrens

Bild: https://owasp.org/www-project-ai-security-and-privacy-guide/

Ein Beispiel: Sicherung der Trainingsdaten

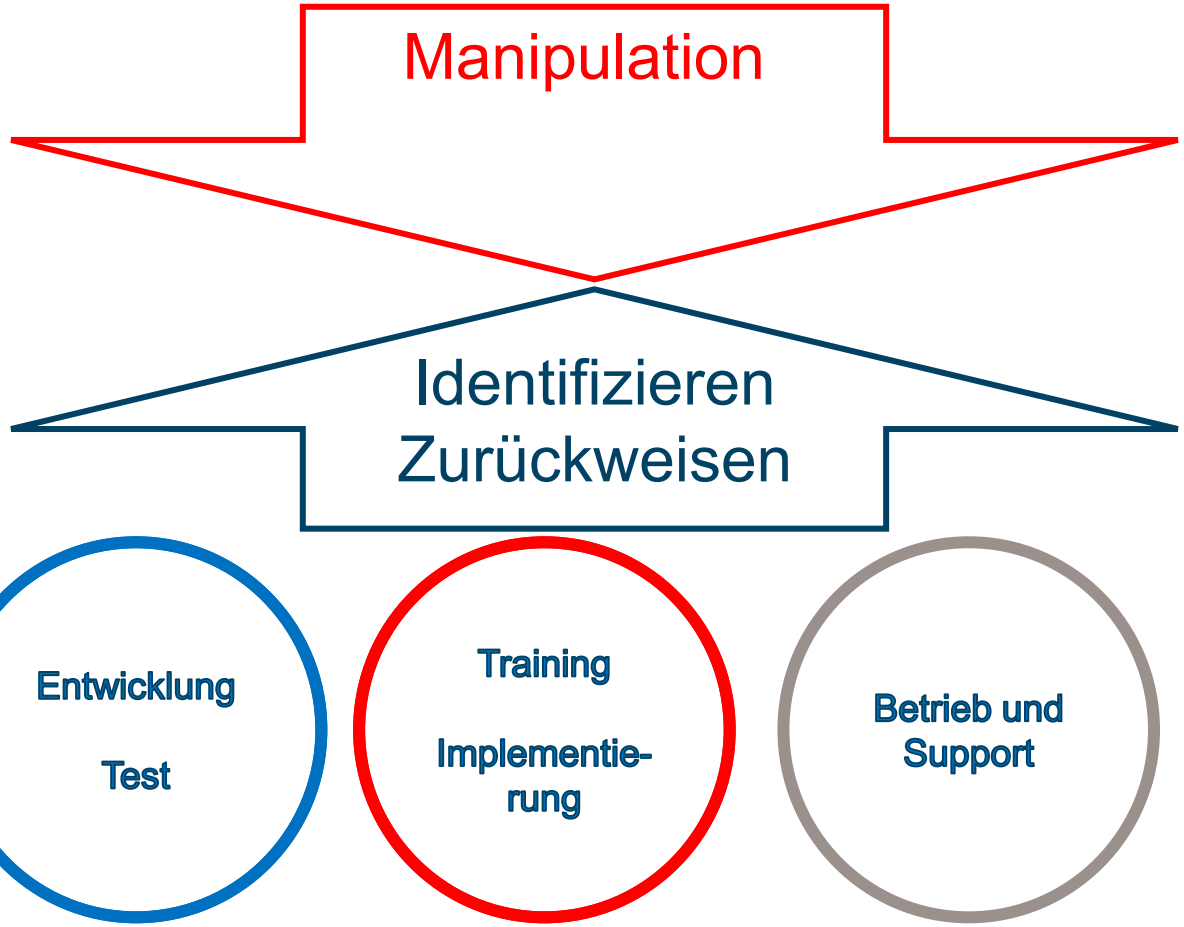
Angriff:
Poisoning (=Manipulation) der Datensätze für das Training eines Machine Learning Verfahrens

Verteidigung mittels „klassischer“ IT-Sicherheit:
Datenintegrität, Datenauthentizität



Ein Beispiel: Sicherung der Trainingsdaten

Angriff:
Poisoning (=Manipulation) der Datensätze für das Training eines Machine Learning Verfahrens
Verteidigung mittels "sicherer" Algorithmen:



Manipulating Machine Learning: Poisoning Attacks and Countermeasures for Regression Learning

Matthew Jagielski*, Alina Oprea*, Battista Biggio ‡‡, Chang Liu†, Cristina Nita-Rotaru*, and Bo Li§

*Northeastern University, Boston, MA †University of Cagliari, Italy ‡Pluribus One, Italy §UC Berkeley, Berkeley, CA

Abstract—As machine learning becomes widely used for automated decisions, attackers have strong incentives to manipulate the results and models generated by machine learning algorithms. In this paper, we perform the first systematic study of poisoning attacks and their countermeasures for linear regression models. In poisoning attacks, attackers deliberately influence the training data to manipulate the results of a predictive model. We propose a theoretically-grounded optimization framework specifically designed for linear regression and demonstrate its effectiveness on a range of datasets and models. We also introduce a fast statistical attack that requires limited knowledge of the training process. Finally, we design a new principled defense method that is highly resilient against all poisoning attacks. We provide formal guarantees about its convergence and an upper bound on the effect of poisoning attacks when the defense is deployed. We evaluate extensively our attacks and defenses on three realistic datasets from health care, loan assessment, and real estate domains.

I. Introduction

As more applications with large societal impact rely on machine learning for automated decisions, several concerns have emerged about potential vulnerabilities introduced by machine learning algorithms. Sophisticated attackers have strong incentives to manipulate the results and models generated by machine learning algorithms to achieve their objectives. For instance, attackers can deliberately influence the training dataset to manipulate the results of a predictive model (in *poisoning* attacks [5], [40]–[42], [45], [48], [56]), cause misclassification of new data in the testing phase (in *evasion attacks* [3], [9], [21], [43], [44], [51], [52]) or infer private information on training data (in *privacy attacks* [19], [20], [46]). Several experts from academia and industry highlighted

training process. Such poisoning attacks have been practically demonstrated in worm signature generation [42], [45], spam filters [40], DoS attack detection [48], PDF malware classification [56], handwritten digit recognition [5], and sentiment analysis [41]. We argue that these attacks become easier to mount today as many machine learning models need to be updated regularly to account for continuously-generated data. Such scenarios require *online training*, in which machine learning models are updated based on new incoming training data. For instance, in cyber-security analytics, new Indicators of Compromise (IoC) rise due to the natural evolution of malicious threats, resulting in updates to machine learning models for threat detection [23]. These IoCs are collected from online platforms like VirusTotal, in which attackers can also submit IoCs of their choice. In personalized medicine, it is envisioned that patient treatment is adjusted in real-time by analyzing information crowdsourced from multiple participants [16]. By controlling a few devices, attackers can submit fake information (e.g., sensor measurements), which is then used for training models applied to a large set of patients. Defending against such poisoning attacks is challenging with current techniques. Methods from robust statistics (e.g., [18], [27]) are resilient against noise but perform poorly on adversarially-poisoned data, and methods for sanitization of training data operate under restrictive adversarial models [13].

One fundamental class of supervised learning is linear regression. Regression is widely used for prediction in many settings (e.g., insurance or loan risk estimation, personalized medicine, market analysis). In a regression task a numerical response variable is predicted using a number of predictor

Jagielski et al. *Manipulating Machine Learning: Poisoning attacks and countermeasures for regression learning* - https://arxiv.org/pdf /1804.00308.pdf - 2021

Ein Beispiel: Sichere Algorithmen

Angriff:
Membership inference attack (durch gezielte Anfragen an das trainierte ML Modell Rückschlüsse auf die Trainingsdaten ziehen)

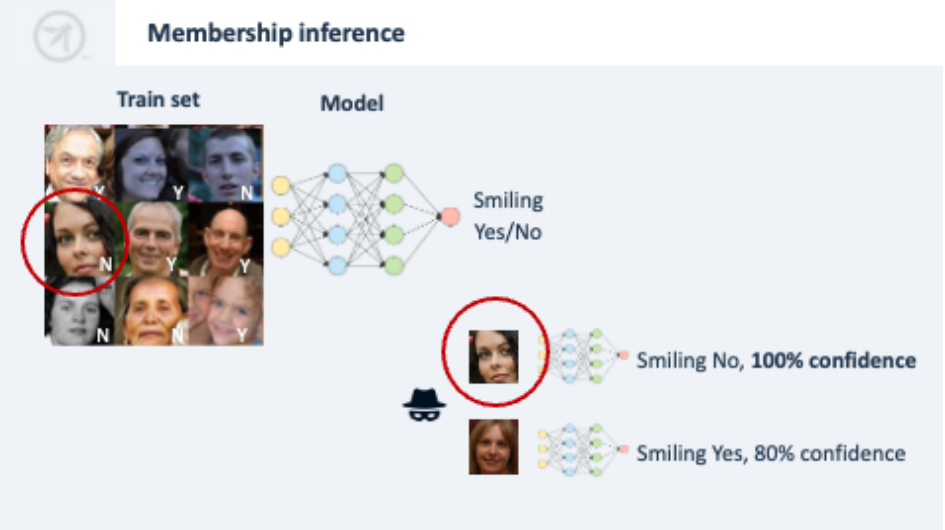


Bild: https://owasp.org/www-project-ai-security-and-privacy-guide/

Ein Beispiel: Sichere Algorithmen

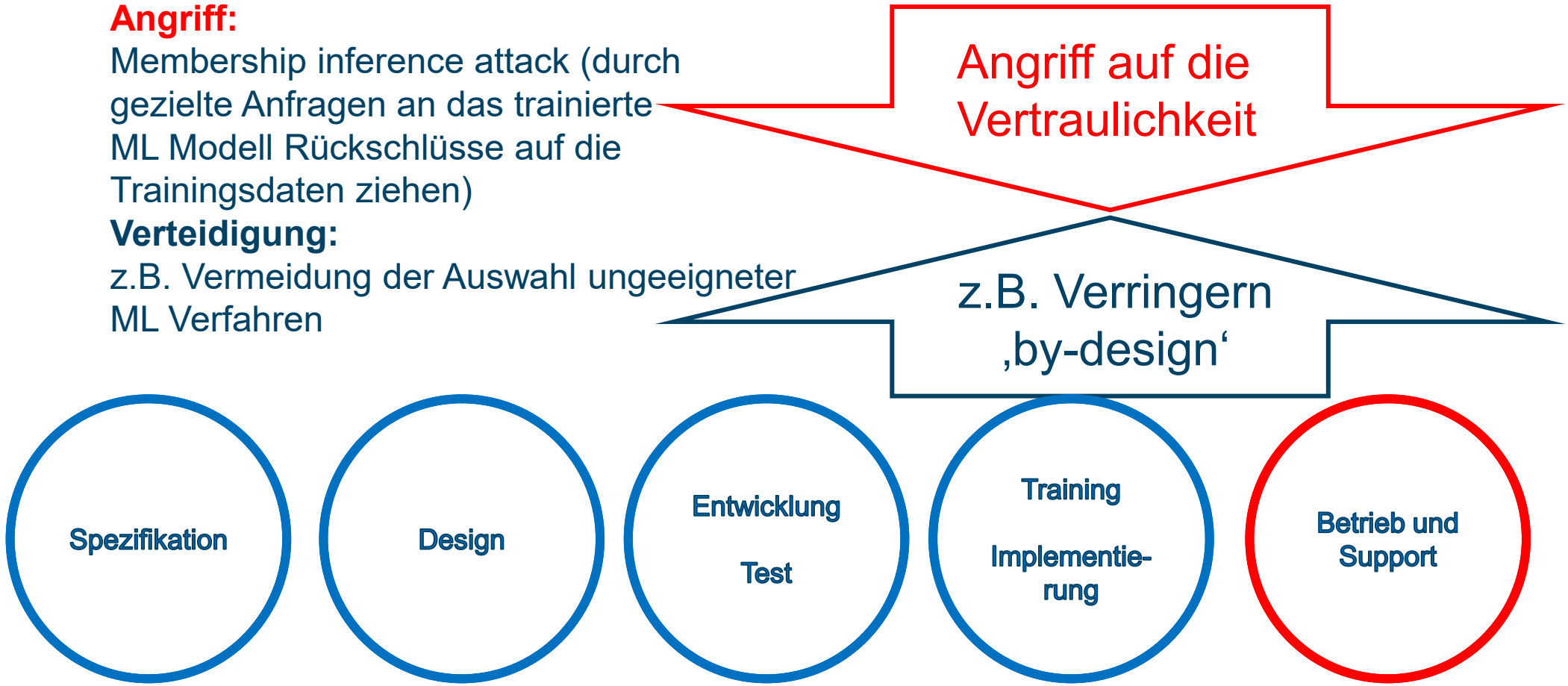
Angriff:
Membership inference attack (durch gezielte Anfragen an das trainierte ML Modell Rückschlüsse auf die Trainingsdaten ziehen)

Angriff auf die Vertraulichkeit

Ein Beispiel: Sichere Algorithmen

Angriff:
Membership inference attack (durch gezielte Anfragen an das trainierte ML Modell Rückschlüsse auf die Trainingsdaten ziehen)

Verteidigung:
z.B. Vermeidung der Auswahl ungeeigneter ML Verfahren



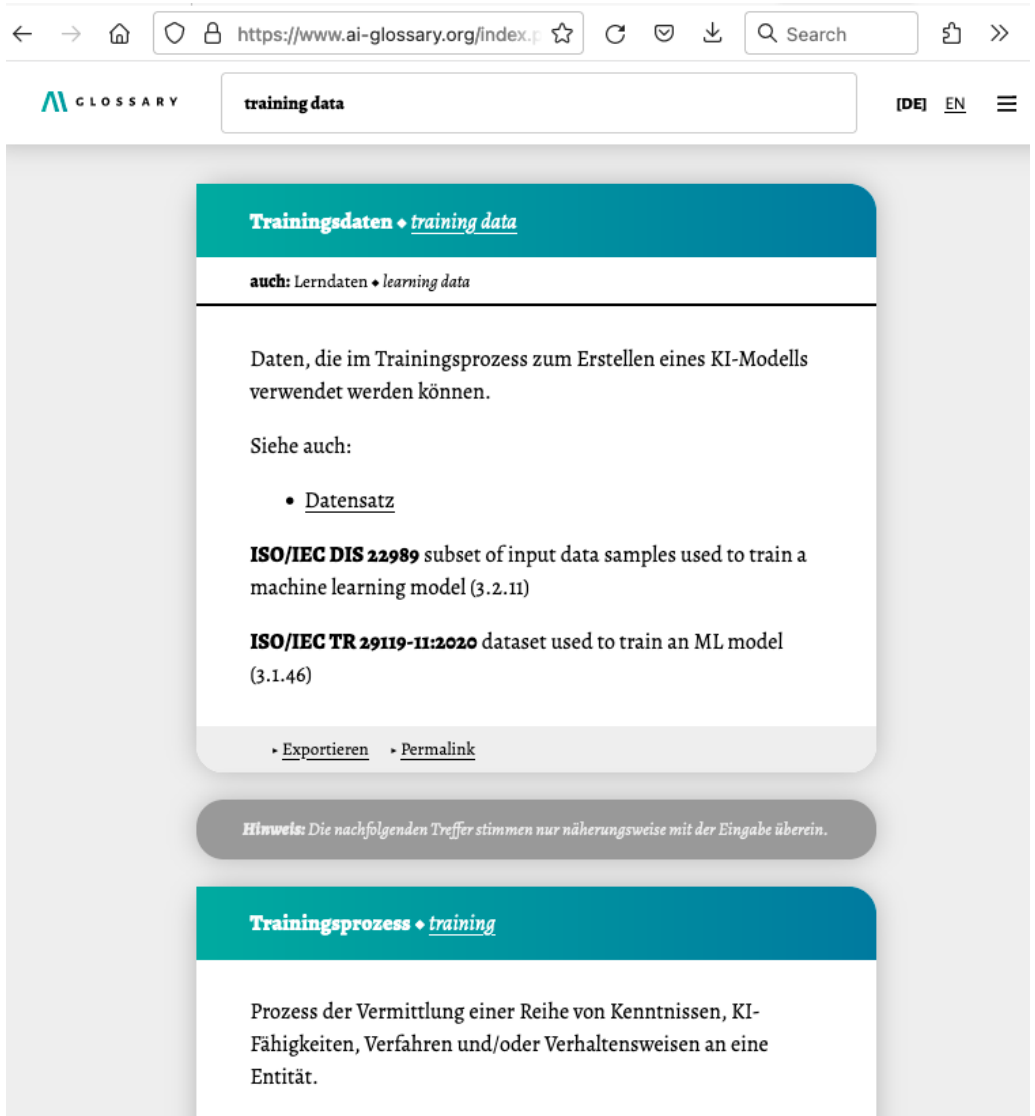
Begriffe für interdisziplinäre Kommunikation

www.ai-glossary.org

Entstanden aus der Glossararbeitsgruppe während der Entwicklung der NRM KI v2

Wird weiter gepflegt und regelmäßig um Begriffe erweitert werden

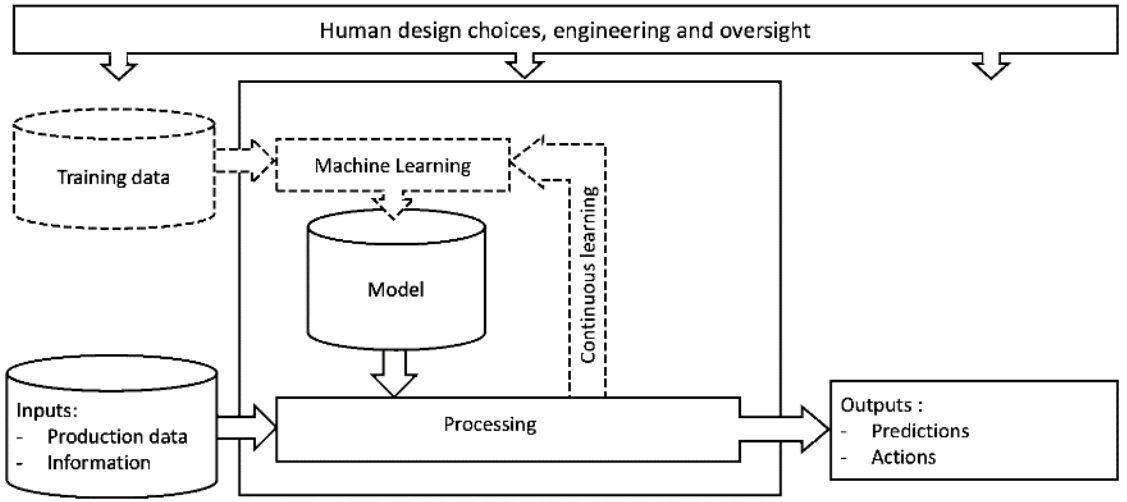
Versionierte Begriffsdefinitionen; Permanente Links; Angabe zu den Standards und Normen



Bedarf 02-05

Abstrakte Zerlegung der KI-Komponente in Daten und Prozesse

- ISO/IEC 22989:2022 macht einen ersten Versuch, dieser ist aber auf sehr hoher Abstraktionsebene.



ISO 22989 – p.41 – Figure 5

- ISO/IEC 27090:draft versucht KI-spezifische Angriffe zu beschreiben, hier könnte ein detaillierteres Architekturdiagram eingebracht werden, um die beschriebenen Angriffe/Risiken hinsichtlich zu lokalisieren

Bedarf 02-06

Existierende KI-Angriffe und Risiken mit existierenden zertifizierbaren IT-Sicherheitszielen abgleichen

- Abbildung von Angriffen auf KI-System als Angriffe auf IT-Sicherheits-Schutzziele entsprechend einer Beschreibung der schutzwürdigen KI-Komponenten und des Lebenszyklus
- Beispiel

  Angriff: “Data Poisoning“

  → Verletztes Schutzziel: Violation integrity of trainingdata until and during „training phase“

  → Control: Integrity, Authenticity (bzw. Data Governance) für Trainigsdaten

  Angriff: „Membership inference attack“

  → Verletztes Schutzziel: Violation of confidentiality of trainingdata during deployment phase
- ISO/IEC 27090:draft versucht KI-spezifische Angriffe zu beschreiben, hier könnte ein erstes Mapping versucht werden
- Ähnlich ISO/IEC 27091:draft für Privacy

Bedarf 02-07

Standardisierung von KI-Produkt- und Prozessprüfverfahren für Security und Privacy

- Wir brauchen für eine Zertifizierung klare Messverfahren und vor allem prüfbare Messwerte für die Security- und Privacy-Eigenschaften von KI-Systemen
- Prüfverfahren und Akkreditierungsverfahren (für die Prüfenden) sind essenziell, um die Qualität der Prüfung durch unabhängige Dritte sicherzustellen
- Passende generische Ansätze
  (ggf. ISMS Ansätze aus der IT-Sicherheit übernehmen)
- Zusätzlich spezielle Prüfungen abhängig von der KI Methode
  oder dem speziellen Algorithmus
- Es gibt erste Ansätze z.B. für Machinelles Lernen
  von der ENISA : https://www.enisa.europa.eu/publications/securing-machine-learning-algorithms

Bedarf 02-08

Ausarbeitung eines horizontalen Querschnittsstandards und vertikale Ausprägungen zu Security

Empfehlenswert wäre die Herausarbeitung von

horizontalen Themen zu Cybersecurity und Privacy für KI zur Prüfung und Zertifizierung, die alle Sektoren betreffen, sowie eine Schnittstelle zu sektorspezifischen Anforderungen.

Ein horizontales Thema wäre beispielsweise die Anforderung an eine geeignete Zugriffskontrolle. Als vertikale Ausprägung können wiederum spezielle Security-Anforderungen aus dem sektoralen Umfeld angesehen werden, wie u. a. für den Bereich der Medizinprodukte.

Themenschwerpunkt: KI-Sicherheit bei IT-Systemen

Bedarf 02-09

Entwicklung von Metriken und Controls gemäß den Standardisierungsanforderungen des geplanten EU AI Act

Entwicklung von Standardisierung zu

Cybersecurity-Anforderungen aus dem AI Act für Metriken und Controls zur Messung und Vermeidung von Cyberangriffen sowie Methoden für Prüfung, Auditierung und Zertifizierung inklusive Anforderungen an die Kriterien für die Prüfmaßnahmen und Prüfenden.

Dabei erscheint es wichtig, eine gemeinsame Arbeitsgruppe mit den Gremien der Cybersecurity und KI in den Standardisierungsorganisationen von Deutschland, der EU und eventuell auch international zu etablieren.

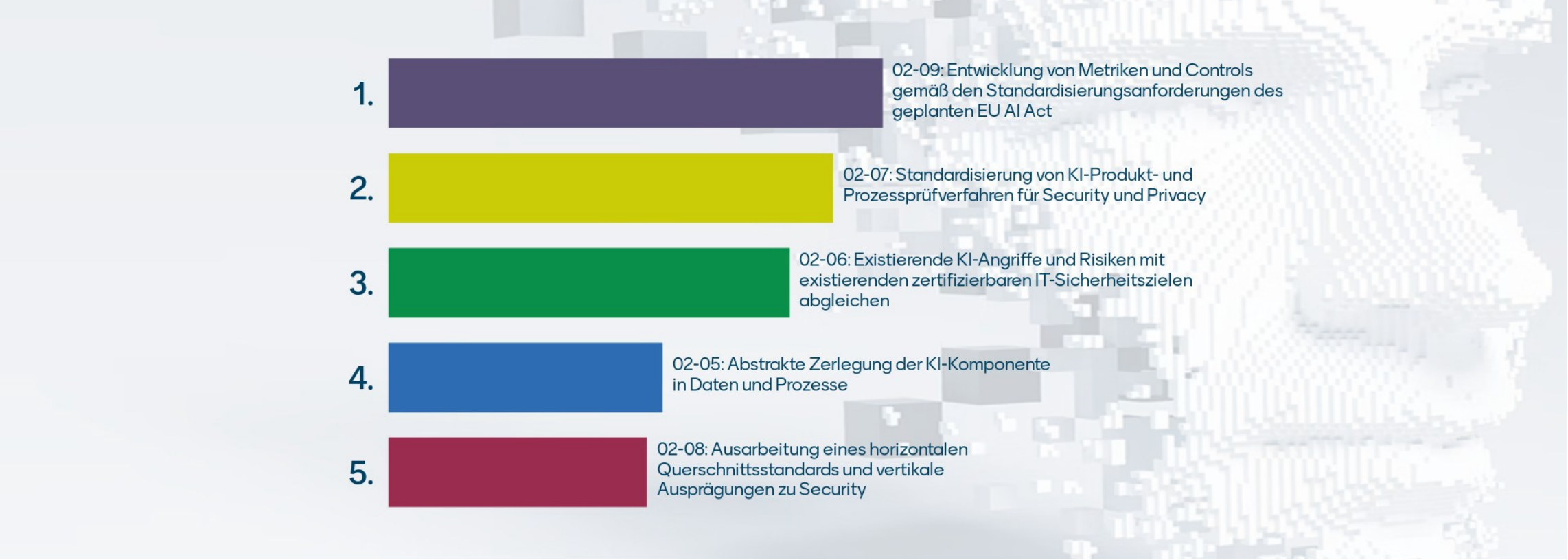
3. Interaktive Priorisierung der Bedarfe (Conceptboard)

Interessierte für die Bedarfsumsetzung

Während des Workshops wurde Conceptboad verwendet. Das Board ist nun geschlossen. Sie können Sich aber noch per E-Mail an Jan.Roesler@din.de für die Umsetzung der Bedarfe melden – bitte Bedarfs-Code mit angeben.

3. Interaktive Priorisierung der Bedarfe (Mentimeter)

Priorisierung der Bedarfe für den Workshop – Ergebnis



Für die Abstimmung konnten alle 5 Bedarfe von 1 bis 5 geordnet werden.

4. Diskussion der Bedarfe anhand der Priorisierung

Blitzlicht zur Diskussion der Bedarfe

- Die Bedarfe 02-09, 02-07 und 02-06 wurden diskutiert und ein gemeinsames Verständnis geschaffen. Allgemein gilt, dass die Bedarfe nicht losgelöst voneinander betrachtet werden können und über die Grenzen der IT-Sicherheit hinaus Schnittmengen mit anderen Bedarfen der Roadmap und anderen Normen/Standards haben. Daher soll verstärkt nach Überscheidungen (wie funktionale Sicherheit, Metriken) gesucht werden.
- Der Bedarf 02-09 erfordert die Entwicklung eines KI-Standards zur Robustness.
- Bedarf 02-07 ist als Ergänzung zu Bedarf 02-09 anzusehen und stellt ein Querschnittsthema dar. Folgende Fragestellungen sind zu beantworten:
  - Welche Standards gibt es hier schon und was ist jetzt (durch KI) anders und muss berücksichtigt werden?
  - Welche Änderungen sind nötig, damit Zertifizierungen inkl. KI möglich werden?
  - Welche Ansätze können aus *Information Security Management System* (ISMS) übernommen werden?
  - Welche Rolle spielt KI-Engineering und andere „KI-Methoden“ mit gezielten Eigenschaften → Cyber Resilliance Act
- Existierende Normen und Standards zu IT-Sicherheit ohne Bezug zu KI bleiben bestehen, müssen aber durch weitere KI-Standards (gerade in Bezug zu KI in IT-Sicherheit) ergänzt werden. Da keine aktueller Standard für IT-Sicherheit auch für KI-Systeme angewandt werden kann. Diese müssen speziell für die Anwendung auf KI-Systeme entwickelt werden.

Aktuelle Anmeldezahlen für die weitere Umsetzung (Nachmeldungen bitte an kuenstliche.intelligenz@din.de)

Bedarf 02-05	**Bedarf 02-06**	**Bedarf 02-07**	**Bedarf 02-08**	**Bedarf 02-09**
Abstrakte Zerlegung der KI-Komponente in Daten und Prozesse	Existierende KI-Angriffe und Risiken mit existierenden zertifizierbaren IT-Sicherheitszielen abgleichen	Standardisierung von KI-Produkt- und Prozessprüfverfahren für Security und Privacy	Ausarbeitung eines horizontalen Querschnittsstandards und vertikale Ausprägungen zu Security	Entwicklung von Metriken und Controls gemäß den Standardisierungsanforderungen des geplanten EU AI Act
8	6	9	8	16

5. Nächste Schritte

Wie geht’s weiter?

DIN
Deutsches Institut für Normung e. V.
Am DIN-Platz
Burggrafenstraße 6
10787 Berlin

www.din.de

Jan Rösler

Projektmanager
DIN Strategische Themenentwicklung KI
Jan.Roesler@DIN.de → Anmeldung für Bedarfsumsetzung