TRUSTIFAI - The AI Certification Hub

# Trusted AI

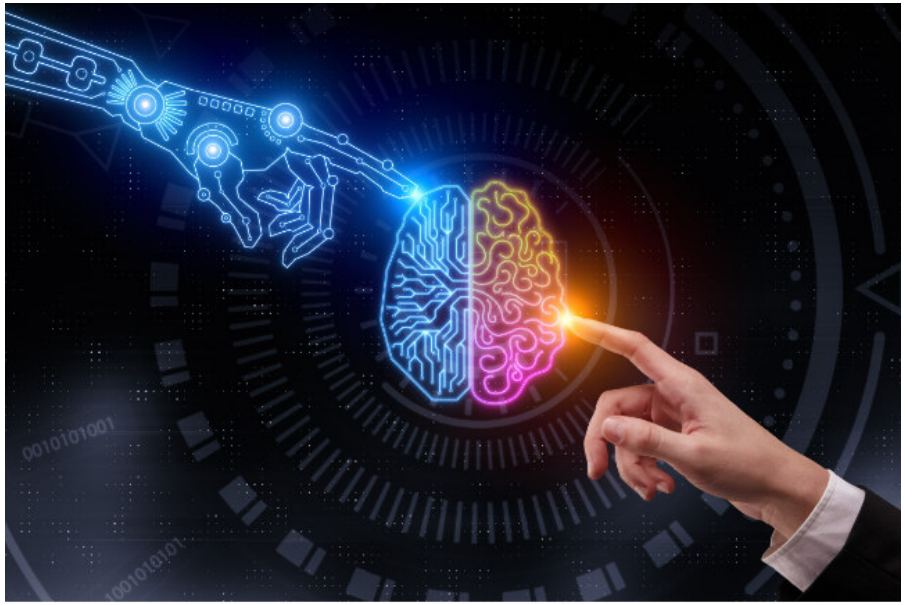
## Prüfung und Zertifizierung von risikobehafteten und sicherheitskritischen *machine learning* Anwendungen

11 October, 2023

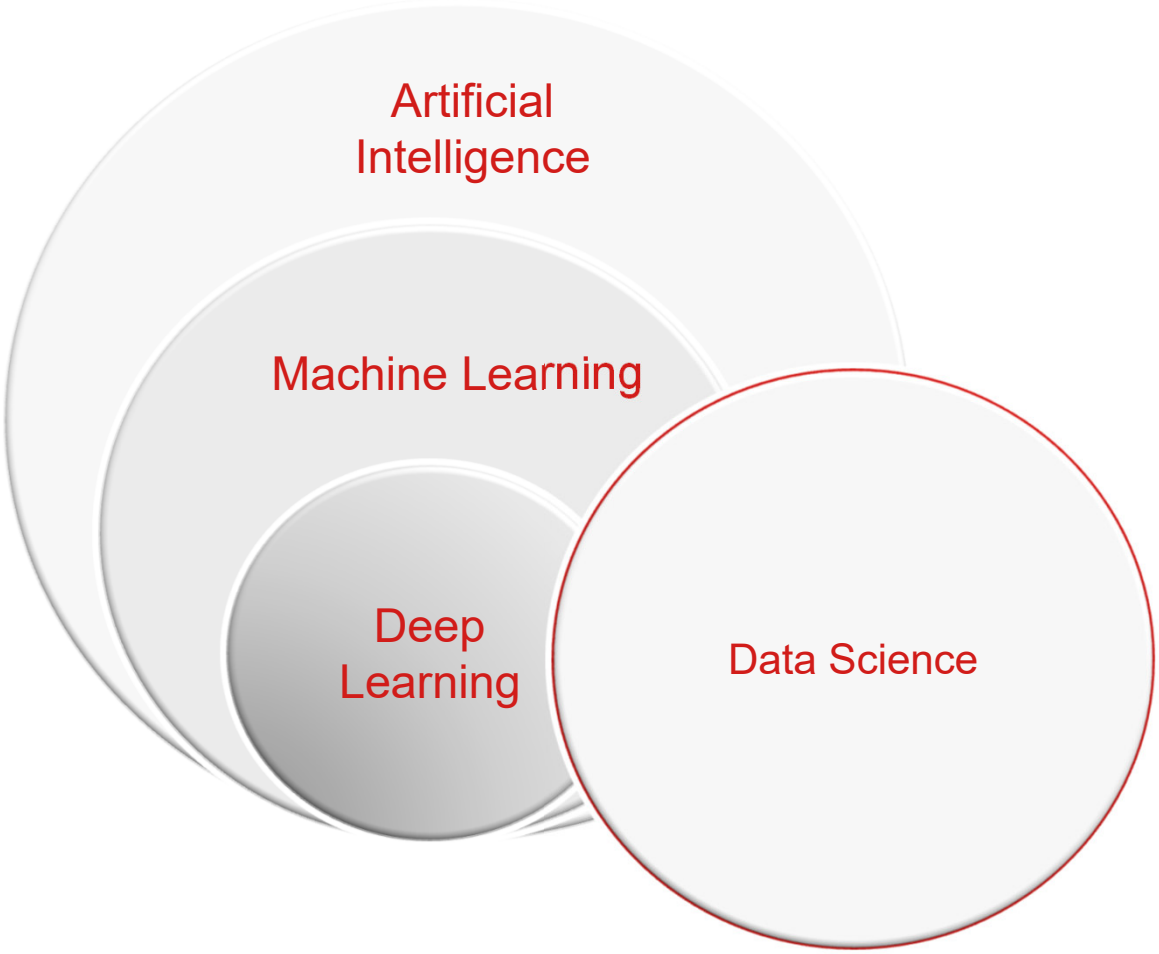
# Was ist AI?

- AI steht für *Artificial Intelligence* – künstliche Intelligenz

- Definition EU AI Act (noch in Diskussion)
  - *‘Artificial intelligence system’ (AI system) means a machine-based system that is designed to operate with varying levels of autonomy and that can, for explicit or implicit objectives, generate output such as predictions, recommendations, or decisions influencing physical or virtual environments.*

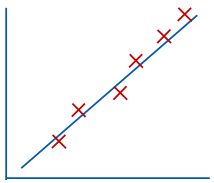


# Einordnung von Artificial Intelligence



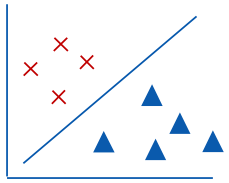
# Klassische Aufgaben des maschinellen Lernens

**„Regression“** – betrachtet statistische Zusammenhänge verschiedener Variablen, um Ergebnisse vorherzusagen



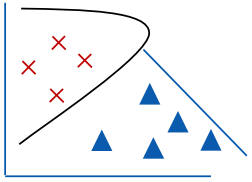
Anwendungen:
Lastprognose für Strom, Fernwärme,
Vorhersage von Ausfällen
etc.

**„Klassifikation“** – unterteilt Daten in verschiedene vorgegebene Klassen



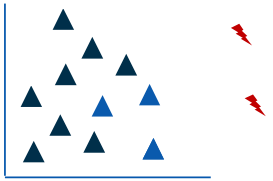
Anwendungen:
Kreditbewertung

**„Cluster-Analyse“** – unterteilt Daten in passende (nicht vorbestimmte) Gruppen



Anwendungen:
Kundensegmentierung

**„Anomalie-Erkennung“** – identifiziert Abweichungen innerhalb einer Datenmenge



Anwendungen:
Betrugserkennung
Erkennung von Messfehlern/ Fehlerhaften Eingaben.

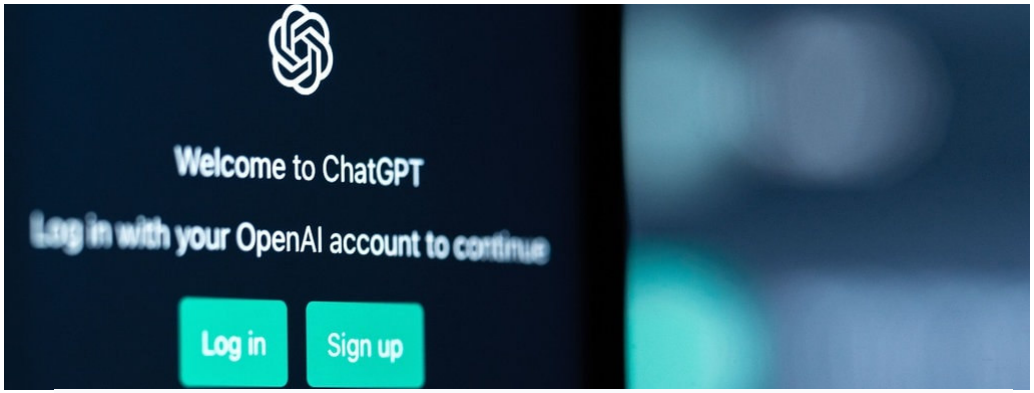
# Generative Modelle

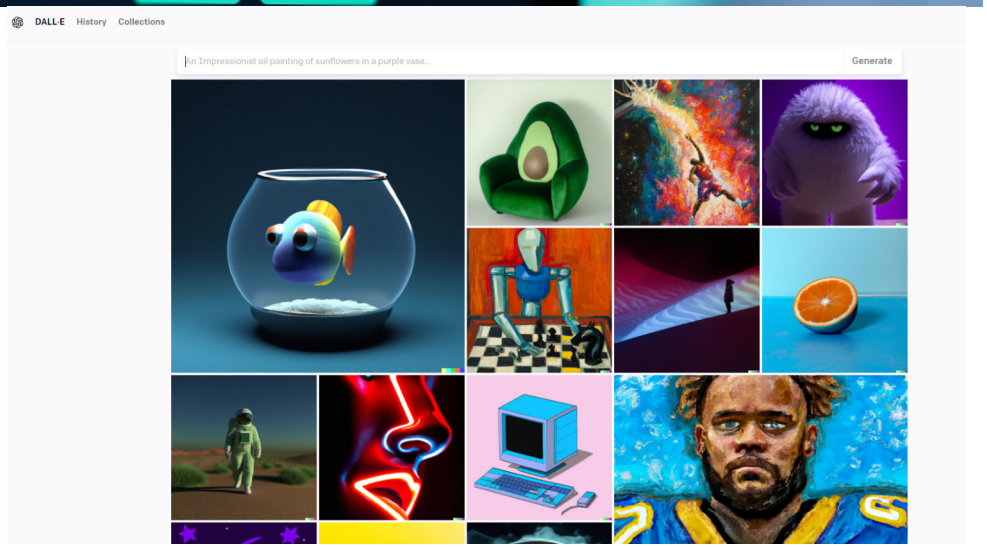
- Generative Modelle können etwas Neues erzeugen

- ChatGPT – generiert Texte
- Stable Diffusion – generiert Bilder
- Code-Vervollständiger – generiert Programmiercode

Offene Fragen:

- Wie sind generative Modelle zertifizierbar?
- Bei wem liegt welche Verantwortung?
- Können generative Modell im sicherheitskritischen Umfeld eingesetzt werden?



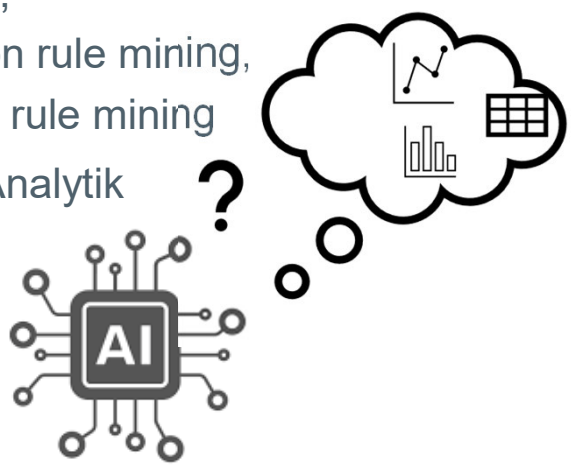



# (Un)Supervised Machine Learning

TRUSTIF AI
by TÜV AUSTRIA Group & scch

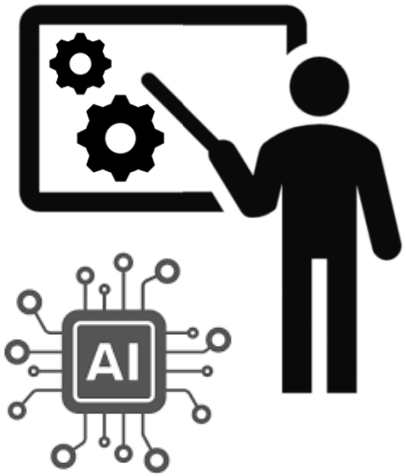
## Unsupervised Learning

- Algorithmus „sucht“ selbst, keine vorgegebene Zielvariable
- Daten werden nach Ähnlichkeit strukturiert
- Dazu zählen z.B:
  - Clustering,
  - Association rule mining,
  - Sequence rule mining
- Deskriptive Analytik




## Supervised Learning

- Zielvariable ist vorgegeben
- Erklärende variable (Prädiktoren beziehen sich auf die Zielvariable)
- Darunter fallen bspw.:
  - Abwanderungsprognose
  - Betrugserkennung
  - Reaktionsmodellierung
  - Kreditrisikomodellierung
- Prädiktive Analytik



# Die Rolle der Zertifizierung

TRUSTIF AI
by TÜV AUSTRIA Group & scch

- ✓ KI ist ein starker Trend in der aktuellen IT, sowohl auf Verbraucher- als auch auf Unternehmensebene.

Wie können wir sicher sein, dass KI in sicherheitskritischen Anwendungen/Systemen wie angekündigt funktioniert?

→ Zertifizierung

- ✓ In Zukunft müssen sicherheitskritische KI-Anwendungen innerhalb eines EU-Rechtsrahmens betrieben werden.
- ✓ Die Zertifizierung von sicherheitskritischer KI muss vorbereitet sein und diese gesetzlichen Anforderungen abdecken.

# Herausforderungen für die Zertifizierung von KI

TRUSTIF AI
by TÜV AUSTRIA Group & scch

- ✓ Aktives, sich ständig veränderndes Feld
- ✓ ... angetrieben durch verbesserte Methoden / Hardware
- ✓ Komplexe theoretische Grundlagen
- ✓ ... mit hohen Eintrittsbarrieren, viel Erfahrung erforderlich
- ✓ Viele verschiedene Methoden des maschinellen Lernens (ML)
- ✓ Viele verschiedene Anwendungen
- ✓ Keine bestehenden Normen

# Herausforderung Standardisierung

- Derzeit gibt es keine international anerkannten Standards für die Zertifizierung von sicheren, zuverlässigen und vertrauenswürdigen KI-Anwendungen.
- Die Normungsgremien arbeiten mit Hochdruck daran, in 1,5 bis 2 Jahren die ersten möglichen Normen herauszubringen.

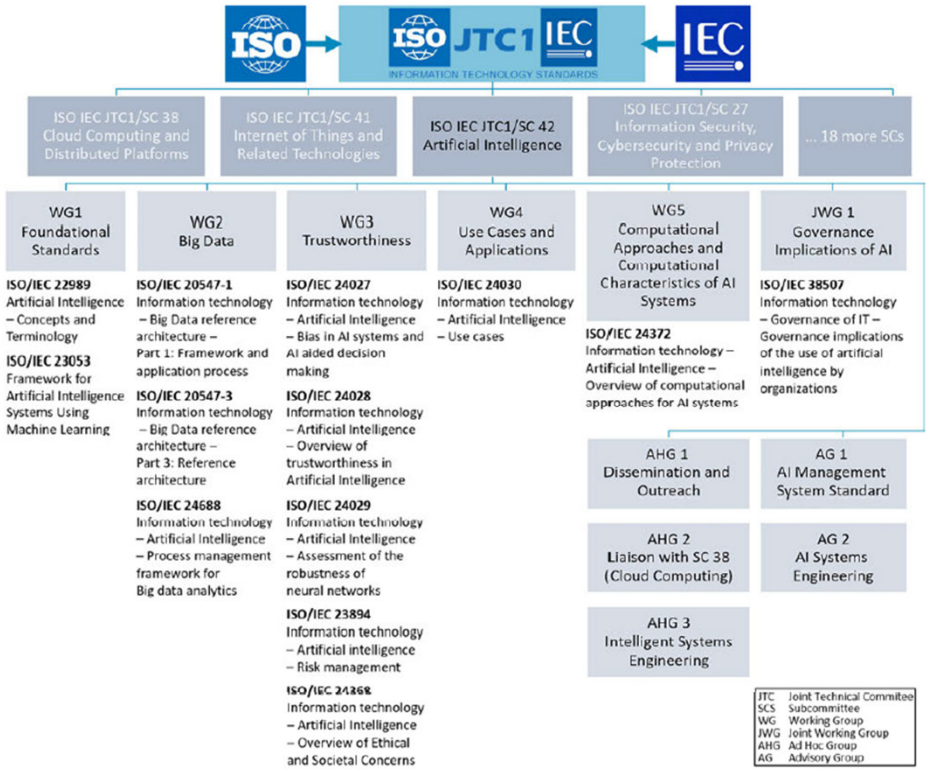


Figure 8: Overview of the work programs and structure of the ISO/IEC JTC 1/SC 42.

# Säulen für die AI-Normung

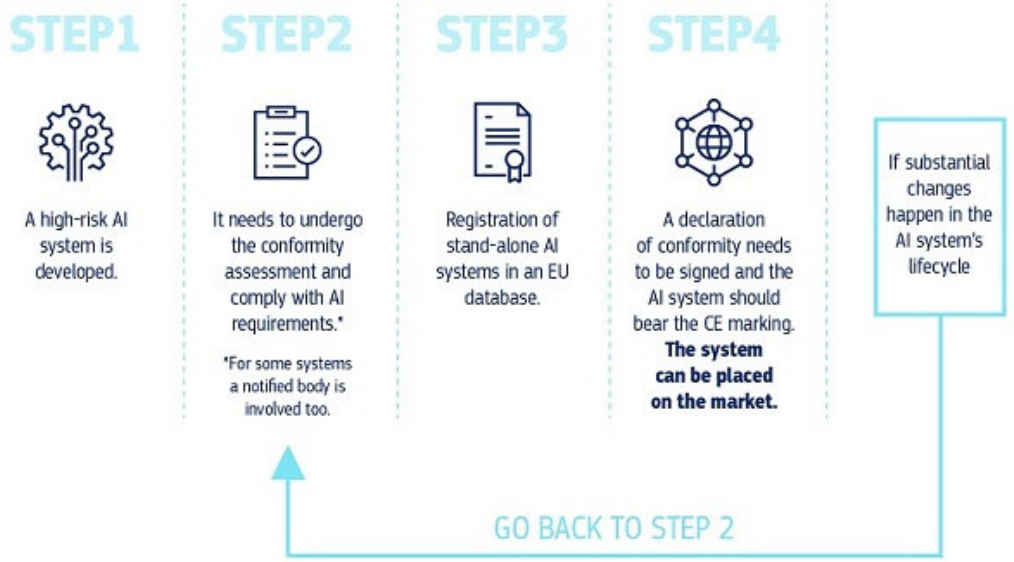
- EC Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL (N0186)
- EU Whitepaper Artificial Intelligence (N0094)
- EU Single Digital Gateway Regulation (N100) und Architecture Vision
- EU Data Strategy (N0093)
- EU Ethic Guidelines for Trustworthy AI (N0041)
- National strategy papers, e.g AIM 2030 – Österreichische KI Strategie (N0069)

# Veröffentlichte Normen

- ISO/IEC TR 24028:2020 Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence
- ISO/IEC TR 20547-5:2018 Information technology — Big data reference architecture — Part 5: Standards roadmap
- ISO/IEC 20547-3:2020 Information technology — Big data reference architecture — Part 3: Reference architecture
- ISO/IEC TR 20547-2:2018 Information technology — Big data reference architecture — Part 2: Use cases and derived requirements
- ISO/IEC TR 20547-1:2020 Information technology — Big data reference architecture — Part 1: Framework and application process
- ISO/IEC 20546:2019 Information technology — Big data — Overview and vocabulary
- ISO/IEC TR 24029-1:2021 Artificial Intelligence (AI) — Assessment of the robustness of neural networks — Part 1: Overview
- ISO/IEC TR 24030:2021 Information technology — Artificial intelligence (AI) — Use cases

# EU-AI Act Risikoklassen und Zeitplan

- 4 verschiedene Risikostufen definiert: minimales Risiko, begrenztes Risiko, hohes Risiko und unannehmbares Risiko. Die Einstufung erfolgt auf der Grundlage des Verwendungszwecks.
- Für Anwendungen mit hohem Risiko müssen Bewertungen durchgeführt werden - die Qualität der Daten, eine angemessene Dokumentation, nachvollziehbare Ergebnisse usw. müssen gewährleistet sein.
- Die Richtlinie über die KI-Haftung wird den EU-Rahmen für die zivilrechtliche Haftung ergänzen und modernisieren, indem sie zum ersten Mal spezifische Regeln für Schäden einführt, die durch KI-Systeme verursacht werden.




- Abstimmung im EU-Parlament in den zuständigen Ausschüssen LIBE und IMCO ist für Donnerstag, 11. Mai 2023, vorgesehen. Bis Ende des Jahres sollen die Verhandlungen mit den Mitglieds-staaten abgeschlossen sein. Insgesamt wurden über 3.000 Abänderungsanträge eingereicht.

# KI Zertifizierung auf Basis des EU AI Act – Situation und Probleme

TRUSTIF AI
by TÜV AUSTRIA Group & scch

Das EU-AI-Act regelt, dass eine große Menge an Dokumentation geschrieben werden muss, aber er versäumt es grob, ein Maß an überprüfbaren Qualitätsanforderungen für zukünftige automatisierte Entscheidungen zu definieren.

Um das zu erkennen, muss man wissen, dass der Unterschied zwischen einem vertrauenswürdigen, professionell entwickelten KI-Entscheidungssystem und einem nicht vertrauenswürdigen, lausigen KI-Entscheidungssystem ausschließlich in der Präzision der Definition seiner Anwendungsdomäne liegt und in der Tatsache, dass es genau in dieser Domäne statistisch getestet wurde.

Unter funktionaler Vertrauenswürdigkeit verstehen wir all eAspekte der Vertrauenswürdigkeit, die direkt von den spezifischen Eigenschaften der ML-Funktion selbst abhängen, die das Ergebnis des datenabhängigen Optimierungs- oder Lernalgorithmus ist. Alle anderen Anforderungen, wie menschliche Aufsicht und Protokollierungspflichten, sind ebenfalls wichtig für vertrauenswürdige KI-Systeme, stehen aber hinter der Vertrauenswürdigkeit der statistischen Qualität der Funktion selbst zurück. Diesen ersten Aspekt nennen wir im Folgenden funktionale Vertrauenswürdigkeit

# Weitere wichtige Themenbereiche für TRUSTED AI

- Model quality
- Quality of performance
- Quality of data
- Quality of development
- Safety, security, privacy
- Transparency
- Conformity
- Drift detection (data drift / model drift)
- Robustness and anomaly detection
- Explainability / Interpretability
- Evaluation of data leakage
- Uncertainty estimation
- Fairness

**Für sicherheitskritische Anwendungen, ist die kntinuierliche Beobachtung des Modells erforderlich!!!!**

# AI-Zertifizierung

TRUSTIF AI
by TÜV AUSTRIA Group & scch

- ✓ Ab 2020 Kooperation zwischen JKU und TÜV AUSTRIA
- ✓ Gestaltung des Trusted AI Audit Katalogs und Veröffentlichung eines Whitepapers

**Johannes Kepler Universität Linz**

JKU LIT Artificial Intelligence Lab

Machine Learning Institute

Univ.-Prof. Dr. Sepp Hochreiter

**Trusted AI audit catalog**
for the certification of ML applications

**TÜV AUSTRIA, Wien**

Next Horizon / DigiTÜV

nexthorizon DIGITÜV
TÜV AUSTRIA Group

# AI-Zertifizierung - Kritikalitätsstufen

TRUSTIF AI
by TÜV AUSTRIA Group & scch

✓ Inspiriert durch die Kritikalitätspyramide der EU-Datenethikkommission

| Kritikalitätsstufe | Wirkungspotenzial (Beispiele) | ML-Anwendungsanforderungen |
|---|---|---|
| 1 | Kein Risiko der Schädigung von Lebewesen, kein Risiko des Verlusts vertraulicher Daten, keine ethischen oder datenschutzrechtlichen Bedenken. | Die grundlegenden Mindestanforderungen an eine kompetent entwickelte ML-Anwendung sind erfüllt. |
| 2 | Lebewesen könnten mit begrenztem, nicht dauerhaftem Schaden geschädigt werden. Vorübergehende Nichtverfügbarkeit von nicht kritischen Daten und Diensten, Verletzung ethischer Bedenken ohne erkennbaren Schaden für Personen. | Die ML-Anwendung wird nach Industriestandards entwickelt und folgt bewährten Verfahren, die als Stand der Technik gelten. |
| 3 | Lebewesen könnten sterben oder in ihrer Lebensfähigkeit eingeschränkt werden; die Umwelt könnte geschädigt werden. Manipulation von Daten mit schwerwiegenden finanziellen Folgen, Verlust der Kontrolle über das System an böswillige Angreifer. Verlust von Informationen, die die Existenz der Organisation gefährden. Langfristige Nichtverfügbarkeit von kritischen Daten oder Diensten, ohne die die Organisation nicht funktionieren kann. Sschwere ethische oder datenschutzrechtliche Bedenken. | Die ML-Anwendung wird mit großer Sorgfalt entwickelt und dokumentiert. Die Sicherheit wird durch Prozesse und Techniken gewährleistet, die über herkömmliche Best Practices und Industriestandards hinausgehen. |
| 4 | Alles, was als eindeutige Bedrohung für die EU-Bürger angesehen wird, soll verboten werden: von Social Scoring durch Regierungen bis hin zu Spielzeug mit Sprachassistenz, das Kinder zu gefährlichem Verhalten anregt. | Inakzeptabel!!!! |

# AI-Zertifizierung - Vertrauenswürdiger AI-Audit-Katalog

TRUSTIF AI
by TÜV AUSTRIA Group & scch

## Sichere Software-Entwicklung

Überprüfung von Entwicklungsmethoden und -umgebung

- Prinzipien der sicheren Softwareentwicklung werden angewendet
- Gewährleistung eines angemessenen Qualitätsniveaus bei der Entwicklung
- Einschließlich Bereitstellung und Wartung (Patches)
- Fokus auf Sicherheit der Anwendungs- und Betriebsumgebung

## Funktionale Anforderungen

Validierung von Daten und Modellen des maschinellen Lernens

- Prüfungskatalog auf der Grundlage theoretischer Grundsätze und bewährter Verfahren im Bereich ML.
- Derzeit anwendbar auf überwachtes und unüberwachtes Lernen, modulare Erweiterungen sind geplant
- Qualitative & quantitative Prüfung durch ML-Experten.

## Ethik und Datenschutz

Berücksichtigung von ethischen und datenschutzrechtlichen Standards

- Ethische Leitlinien der EU für vertrauenswürdige KI" (falls zutreffend).
- Konformität mit GDPR (falls zutreffend)

# AI-Zertifizierung – Vertrauenswürdiger AI-Audit-Katalog

TRUSTIF AI
by TÜV AUSTRIA Group & scch

- ✓ Der Prüfungskatalog gilt derzeit für supervised und unsupervised ML-Systeme mit geringem und mittlerem Risiko, Kriterien für Systeme mit hohem Risiko sind derzeit in der Implementierung.
- ✓ Drei Ha … n.
- ✓ Besond …

Aus Sicht der Funktionsprüfung von KI-Anwendungen sind drei Elemente notwendig, um eine zuverlässige funktionale Vertrauenswürdigkeit herzustellen, nämlich (1) die Definition der technischen Verteilung der Anwendung, (2) die risikobasierten Mindestleistungsanforderungen und (3) die statistisch validen Tests auf der Grundlage unabhängiger Stichproben.

Bedenken hinsichtlich Regelungen im EU AI Act:

- Fixe Testsets sind unzureichend
- Fehlende Festlegung von Anwendungsdomänen
- Regulierung der Zusammensetzung von Trainingsdatensets ist nutzlos

| | |
|---|---|
| 1) ML … | |
| 2) De … | |
| 3) Da … | … it |
| 4) Da … | |
| 5) Vo … | |
| 6) Modellentwicklung | 12) Kommunikation |

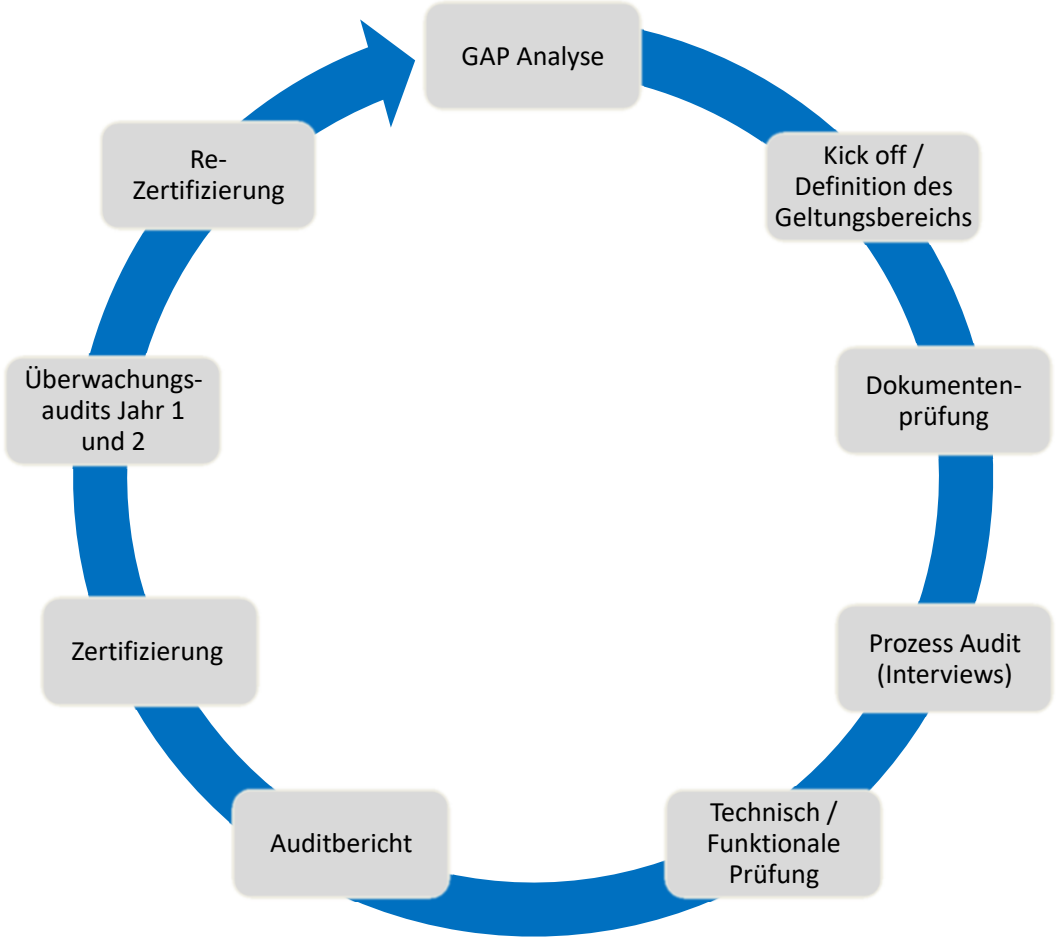
# AI-Zertifizierung – Vertrauenswürdiger AI-Audit-Katalog

TRUSTIF AI
by TÜV AUSTRIA Group & scch

## Secure Software Development

| | |
|---|---|
| Awareness | Accountability |
| Safety and Business Impact - Risk Assessment | Implementation |
| Training | Security Testing |
| Specification | Deployment |
| Design | Patch management |
| Concepts | Quality and integrity of data |
| Technical Procedures | Security response |
| Environment | Security metrics |
| Cloud and third party sources | Agility |
| Technical stability and security | Container |

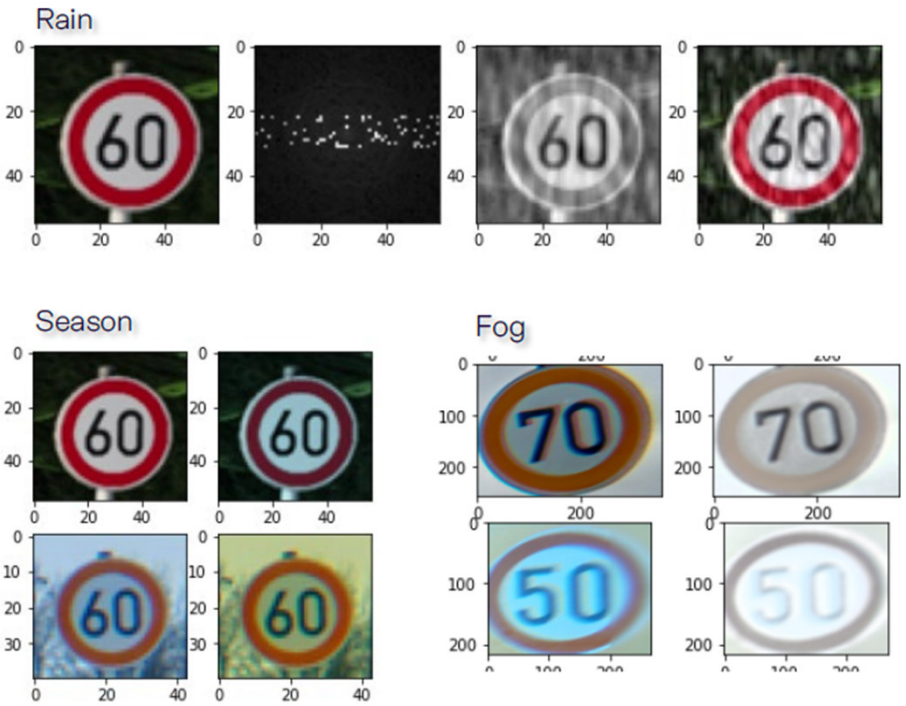
# Audit Prozess



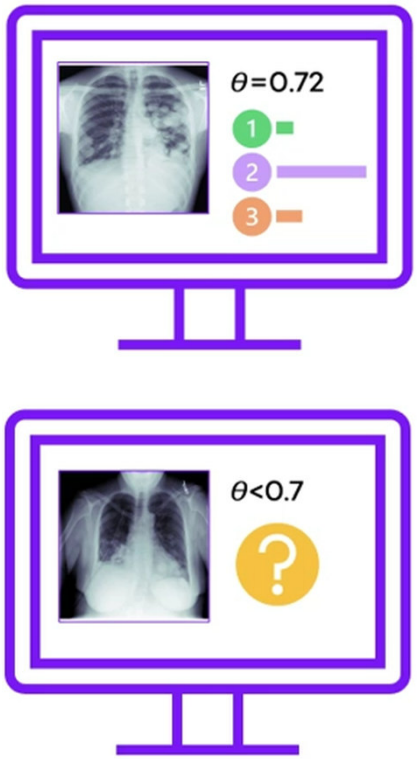
# AI:TIC - Was, wenn... das KI-System versagt?

TRUSTIF AI
by TÜV AUSTRIA Group & scch

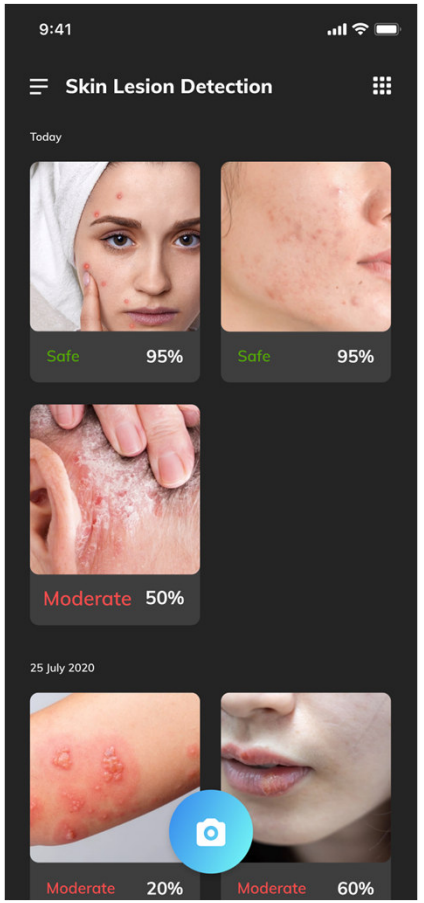
Risiko: Anreicherung mit synthetischen Daten als Ausdruck der realen Welt & gegnerischer Angriffe



# Typische Prüfungsergebnisse im Gesundheitswesen




Problem:
Genauigkeit der Klassifizierung und hoher Grad an Interpretierbarkeit und Erklärbarkeit der Systementscheidungen




Starke Datenverzerrung: nur helle Hauttöne im Trainingssatz enthalten

# Abgrenzung der MVO zum AI- und Cyber Resilience Act

- Der 2022 noch enthaltene Artikel mit Bezug auf den AI-Act wurde nicht in die finale Maschinenverordnung übernommen. Der Text „self-evolving behavoiur“ taucht dennoch u. a. im Annex I Part A in Verbindung mit Sicherheitsbauteilen und Annex III 1.1.6 zur Ergonomie auf.
- Maschinen sind je nach Definition im AI-Act enthalten.
- Eine Maschine fällt unter den AI-Act, wenn sie ein Sicherheitsbauteil ist und Drittstellenpflichtig oder die AI selber eine Maschine ist; z. B. Co-Bots − Die Definition von AI ist noch nicht sicher. Derzeit ist die Definition von AI nicht „Software“ sondern „ein System“ und somit kann es sehr viel sein.
- Abgrenzung Maschinenverordnung zum Cyber Resilience Act (CRA) − der CRA behandelt weniger Safety als vielmehr Security.
- Hat man als Hersteller ein unter die Maschinenverordnung fallendes Produkt, das die Anforderungen aus Anhang III 1.1.9 und / oder 1.2.1 zu erfüllen hat, wird man diese Anforderungen mit dem CRA abdecken müssen.
- Auch wenn AI-Act und CRA verabschiedet wurden, wird erst über den Leitfaden zur Maschinenverordnung mehr Aufklärung erwartbar sein.

# Erweiterung des Fehlerbegriffs in neuer Produkthaftungs-Richtlinie

*Artikel 6*

**Fehlerhaftigkeit**

1. Ein Produkt gilt als fehlerhaft, wenn es nicht die Sicherheit bietet, die die breite Öffentlichkeit unter Berücksichtigung aller Umstände, insbesondere der nachfolgenden, erwarten darf:

   a) der Aufmachung des Produkts, einschließlich der Anweisungen für Installation, Verwendung und Wartung;

   b) der vernünftigerweise vorhersehbaren Nutzung und missbräuchlichen Nutzung des Produkts;

   c) der Auswirkungen einer etwaigen Fähigkeit, nach Einsatzbeginn weiter zu lernen, auf das Produkt;

   d) der Auswirkungen anderer Produkte auf das Produkt, bei denen nach vernünftigem Ermessen davon ausgegangen werden kann, dass sie zusammen mit dem Produkt verwendet werden;

   e) des Zeitpunktes, zu dem das Produkt in Verkehr gebracht oder in Betrieb genommen wurde, oder, wenn der Hersteller nach diesem Zeitpunkt die Kontrolle über das Produkt behält, des Zeitpunktes, ab dem das Produkt nicht mehr unter Kontrolle des Herstellers steht;

   f) der Sicherheitsanforderungen des Produkts einschließlich sicherheitsrelevanter Cybersicherheitsanforderungen;

   g) Eingriffe einer Regulierungsbehörde oder eines in Artikel 7 genannten Wirtschaftsakteurs im Zusammenhang mit der Produktsicherheit;

   h) der spezifischen Erwartungen der Endnutzer, für die das Produkt bestimmt ist.

2. Ein Produkt gilt nicht allein deshalb als fehlerhaft, weil ein besseres Produkt, einschließlich Aktualisierungen oder Upgrades eines Produkts, bereits in Verkehr oder in Betrieb ist bzw. künftig in Verkehr gebracht oder in Betrieb genommen wird.

- Ein Produkt ist fehlerhaft, wenn es nicht der berechtigten Sicherheitserwartung der breiten Öffentlichkeit entspricht.
- Dabei sind nun ausdrücklich zu berücksichtigen:
  - Lernende Systeme („KI“)
  - Kombinationsrisiken
  - Spätere Updates des Produkts
  - Cybersecurity (vgl. auch Typgenehmigung bzw. UN/ECE Nr. 155, 156)
- Außerdem: Verzahnung mit dem Produktsicherheitsrecht.

# ISO/IEC TR 5469 “Artificial Intelligence – Functional safety and AI systems”

- Die ISO/IEC TR 5469 beschreibt Eigenschaften, verbundene Risiken, verfügbare Methoden und Prozesse zur Nutzung der künstlichen Intelligenz (KI):
  - bei Verwendung innerhalb einer sicherheitsrelevanten Funktion zur Realisierung ihrer Funktionalität.
  - bei der sicherheitsrelevanten Funktionen ohne KI, die aber Systeme mit KI absichern sollen. (Beispiel EUC als KI).
  - bei Verwendung von KI-Systemen zum Entwurf und Entwicklung von sicherheitsrelevanten Funktionen.

# Was umfasst die ISO/IEC TR 5469?

TRUSTIF AI
by TÜV AUSTRIA Group & scch

- Eine Übersicht von Methoden und Entwicklungsprozessen die unter den Aspekten der funktionalen Sicherheit kritisch sein können. Teilweise von konventionellen Entwicklungsmethoden der funktionalen Sicherheit abweichen und entsprechend analysiert werden müssen.
- Eine Klassifzierungsmethode für das Verhältnis zwischen Anwendungsbereich und verwendbarer KI-Technologie.
- Eine Verbindung zwischen KI Methoden und Prozessen zu den systematischen Fehlerreduktionsmethoden der IEC 61508 (Technique or Measure).

# Was umfasst die ISO/IEC TR 5469?

TRUSTIF AI
by TÜV AUSTRIA Group & scch

- Eine Übersicht von Methoden und Entwicklungsprozessen die unter den Aspekten der funktionalen Sicherheit kritisch sein können. Teilweise von konventionellen Entwicklungsmethoden der funktionalen Sicherheit abweichen und entsprechend analysiert werden müssen.
- Eine Klassifzierungsmethode für das Verhältnis zwischen Anwendungsbereich und verwendbarer KI-Technologie.
- Eine Verbindung zwischen KI Methoden und Prozessen zu den systematischen Fehlerreduktionsmethoden der IEC 61508 (Technique or Measure).

Was dieser TR nicht liefert!

- Anforderungen an KI Technologie für die funktionale Sicherheit.
- Zusammenhänge zu einem SIL oder einen SC.
- Aussagen über die Wirksamkeit von Maßnahmen.

# Wichtige Definitionen

- “risk, functional safety risk”: Wie in IEC 61508-4, eine Kombination aus der Wahrscheinlichkeit des Auftretens einer Gefahr und dessen Schwere.
- “risk, organizational risk”: Wie in ISO 31000, Auswirkung von Unsicherheit auf die Ziele. Sie kann positive, negative oder beides sein und Chancen und Gefahren ansprechen, schaffen oder zu ihnen führen.
- “AI technology”: Verwendete Technologie zur Implementierung von KI Modellen.
- “AI model”: physikalische, mathematische oder andere logische Darstellung eines Systems, Einheit, Phänomen, Prozesses oder Daten.

# KI Technologieklassen

TRUSTIF AI
by TÜV AUSTRIA Group & scch

| KI Technologieklasse | Definition |
|---|---|
| I | Wenn die KI-Technologie unter Verwendung bestehender internationaler Normen zur funktionalen Sicherheit entwickelt und überprüft werden kann. |
| II | Wenn die KI-Technologie nicht vollständig anhand der bestehenden internationalen Normen für funktionale Sicherheit entwickelt und geprüft werden kann, es aber dennoch möglich ist, die gewünschten Eigenschaften und die Mittel zu ihrer Erreichung durch eine Reihe von Methoden und Techniken zu ermitteln. |
| III | Wenn die KI-Technologie nicht auf der Grundlage bestehender internationaler Normen für funktionale Sicherheit entwickelt und geprüft werden kann und es auch nicht möglich ist, eine Reihe von Eigenschaften mit entsprechenden Methoden und Techniken zu deren Erreichung zu ermitteln. |

# Verwendungsebenen

| Verwendungsebene, Usage level | Definitionen |
|---|---|
| **A1** | Wenn die KI-Technologie in einem funktionalen sicherheitsrelevanten E/E/PE-System eingesetzt wird und eine automatisierte Entscheidungsfindung der Systemfunktion mit Hilfe der KI-Technologie möglich ist. |
| **A2** | Wenn die KI-Technologie in einem sicherheitsrelevanten E/E/PE-System eingesetzt wird und keine automatisierte Entscheidungsfindung der Systemfunktion mit Hilfe der KI-Technologie möglich ist (z. B. wenn die KI-Technologie für Diagnosefunktionen innerhalb des E/E/PE-Systems eingesetzt wird). |
| **B1** | Wenn die KI-Technologie nur während der Entwicklung des sicherheitsrelevanten E/E/PE-Systems eingesetzt wird (z. B. ein Offline-Unterstützungstool) und wenn eine automatisierte Entscheidungsfindung der mit Hilfe der KI-Technologie entwickelten Funktion möglich ist. |
| **B2** | Wenn die KI-Technologie nur während der Entwicklung des sicherheitsrelevanten E/E/PE-Systems eingesetzt wird (z. B. ein Offline-Unterstützungstool) und keine automatisierte Entscheidungsfindung der Funktion möglich ist. |
| **C** | Wenn die KI-Technologie nicht Teil einer funktionalen Sicherheitsfunktion im E/E/PE-System ist, aber indirekt Einfluss auf die Funktion haben kann. |
| **D** | Wenn die KI-Technologie nicht Teil einer Sicherheitsfunktion im E/E/PE-System ist und aufgrund einer ausreichenden Trennung und Verhaltenskontrolle keinen Einfluss auf die Sicherheitsfunktion hat. |

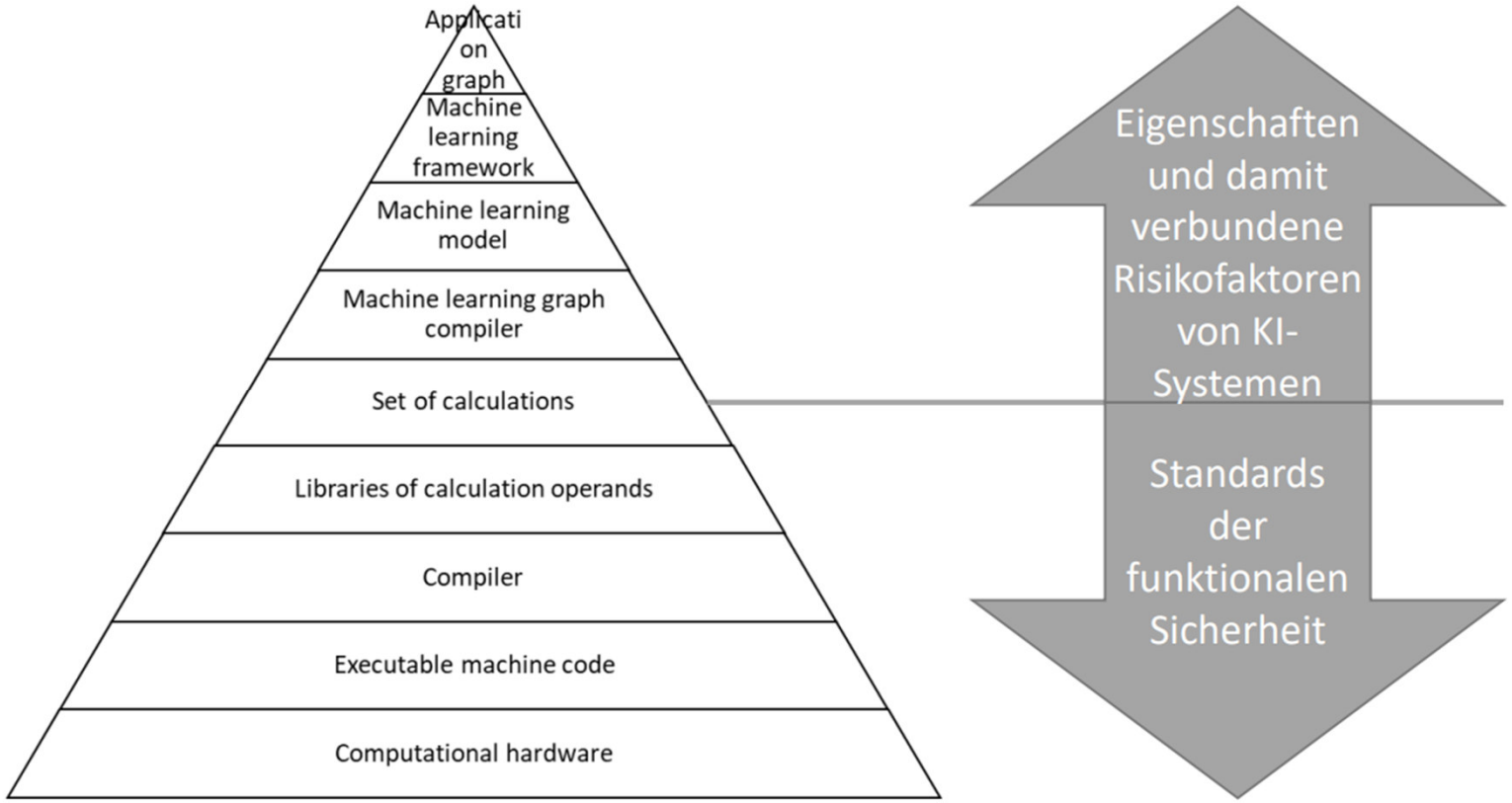
# KI Klassifikationstabelle

TRUSTIF AI
by TÜV AUSTRIA Group & scch

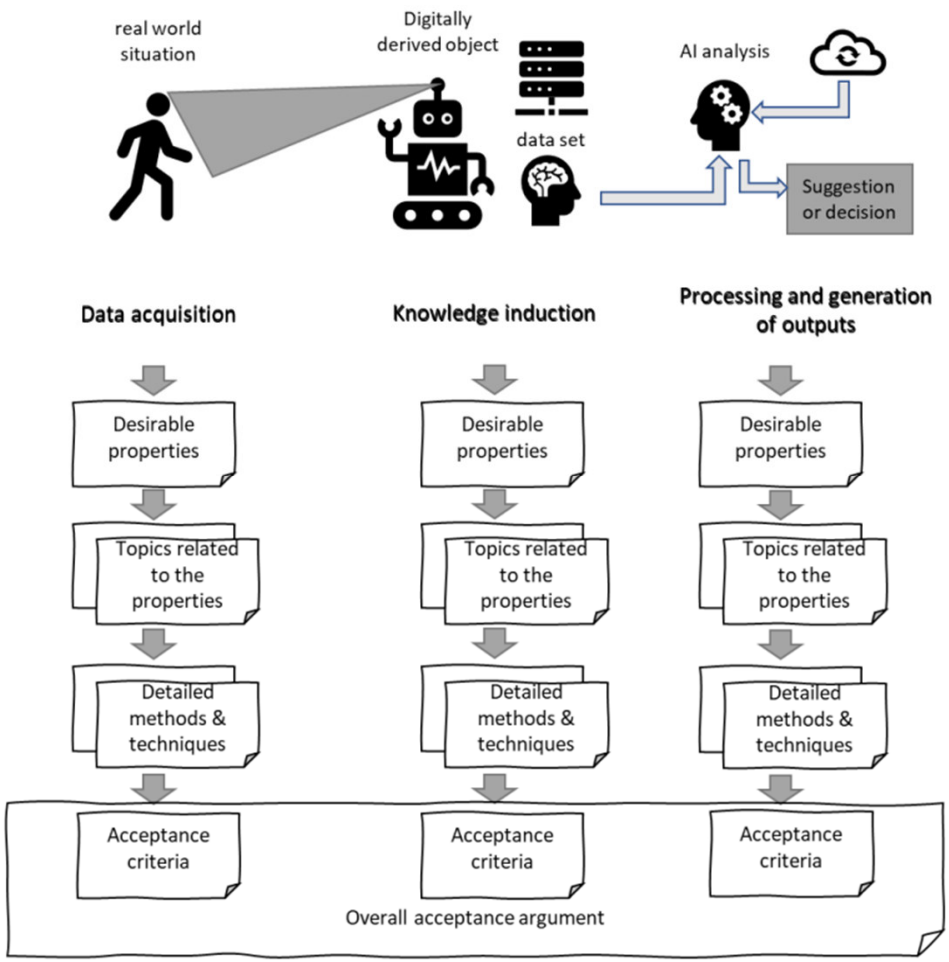
| AI application usage level | AI technology Class I | AI technology Class II | AI technology Class III |
|---|---|---|---|
| **A1** (statisch, nur Offline-Training) | Anwendung der Risiko-minderungskonzepte bestehender internationaler Normen zur funktionalen Sicherheit möglich. | Der geeignete Satz von Anforderungen für jede Nutzungsstufe kann unter Berücksichtigung der Abschnitte 8, 9, 10 und 11 festgelegt des ISO IEC TR 5469 werden. | Nicht empfohlen |
| **A2** (statisch, nur Offline-Training) | | | |
| **B1** (statisch, nur Offline-Training) | | | |
| **B2** (statisch, nur Offline-Training) | | | |
| **C** (statisch, nur Offline-Training; KI liefert ergänzende Risikoreduzierung) | | | |
| **D** (dynamisch, Online-Training ist möglich) | Keine spezifischen Anforderungen an die funktionale Sicherheit von KI-Technologien, aber Anwendung der Risikominderungskonzepte bestehender internationaler Normen zur funktionalen Sicherheit. Darüber hinaus können weitere Sicherheitsaspekte durch den Einsatz von KI möglicherweise beeinträchtigt werden. | | |

- ☑ Der Nutzungsgrad für sichere Anwendungen muss jedoch im Voraus festgelegt werden.
- ☑ Und diese Nutzungsgrade müssen mit verschiedenen KI-Anforderungen verknüpft werden, was noch nicht wirklich definiert und gelöst ist.

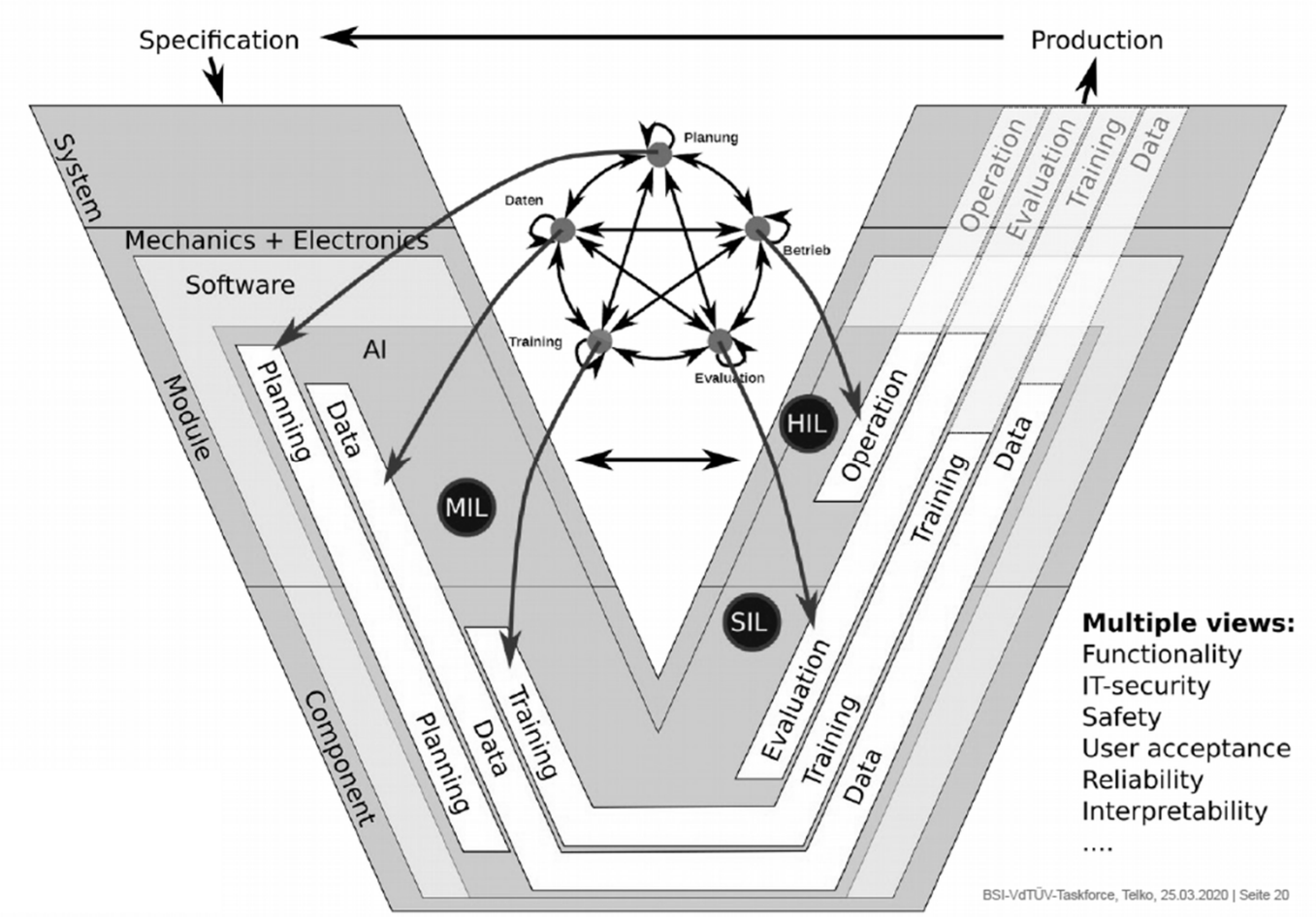
# Hierarchie von Technologieelementen



# KI Entwicklungsprozess



# Mapping of functional requirements



# Eigenschaften und damit verbundene Risikofaktoren von KI-Systemen

TRUSTIF AI
by TÜV AUSTRIA Group & scch

- Grad der Automatisierung und Steuerung. Abhängigkeit einer menschlichen Überwachung.
- Grad der Transparenz und der Erklärbarkeit von KI Systemen.
- Komplexität und unbekannte Vorgaben des Einsatzfeldes.
- Widerstandsfähigkeit gegenüber manipulativen Einflüssen.
- KI spezifische Hardwareprobleme.
- Reifegrad der Technologie.

# Verifikations- und Validierungstechniken

TRUSTIF AI
by TÜV AUSTRIA Group & scch

- Es gibt noch genügend Probleme im Zusammenhang mit der Verifizierung und Validierung
- Mögliche Methoden zur Verifikation und Validation von KI Technologien wie Functional Trustworthiness sind noch nicht spezifiziert undteilweise noch Thema der Grundlagenforschung
- Virtuelles und reales Testen
- Überwachung und Rückmeldung von Vorfällen

# Fehlerbeherrschung und -vermeidung

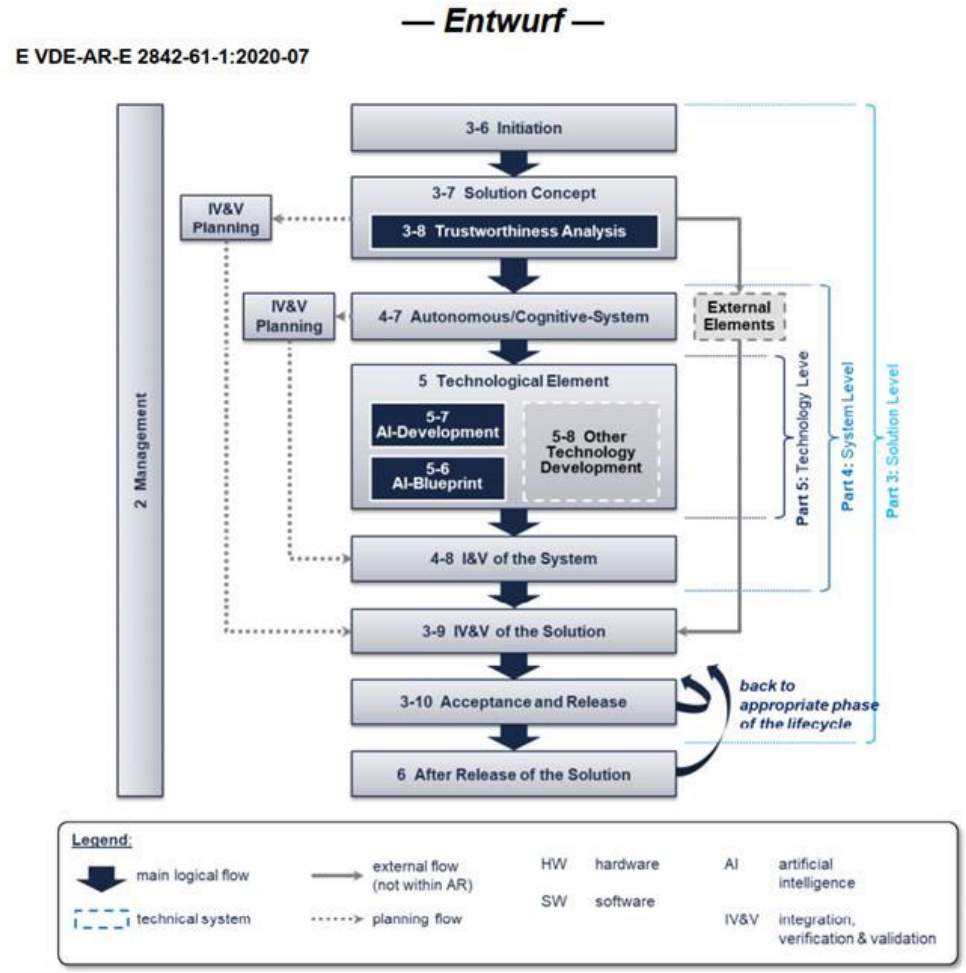
- Was benötigen wir noch:
- Architektur-Betrachtungen
- Erhöhung der Zuverlässigkeit von Komponenten mit KI-Technologie.
- Beziehung zwischen dem AI-Lebenszyklus und dem Lebenszyklus der funktionalen Sicherheit
- Beziehung zwischen den Tabellen der IEC 61508-3 mit der Interpretation für KI Technologieelemente.

# Annex B und Annex C der ISO/IEC TR 5469

enthalten Beispiele

- Für die Anwendung des “Drei-Stufen Realisierungsprinzip!”
- Mögliche Verfahren und nützliche Technologien für die Verifizierung und Validierung

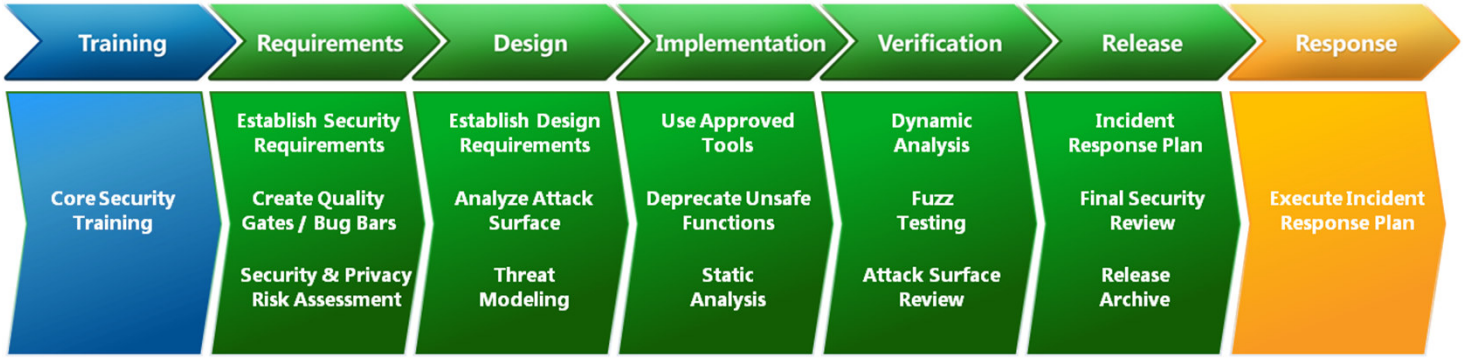
# Weitere Ansätze und Herausforderungen



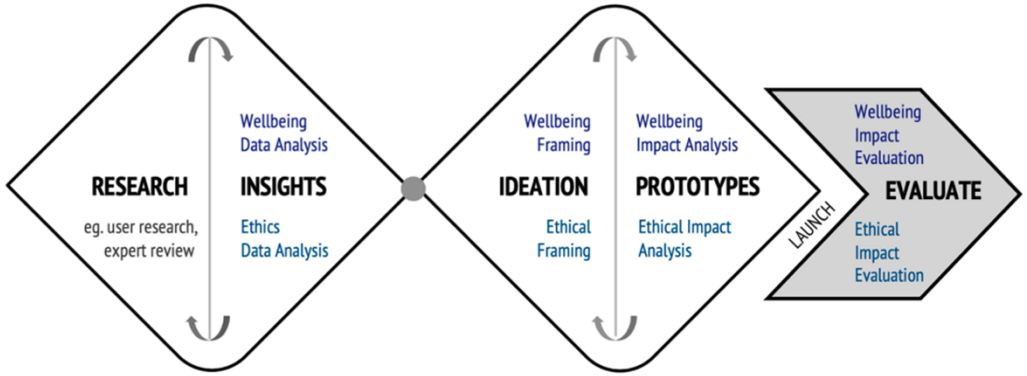

- Alle Ansätze befinden sich noch im Entwicklungsstadium.
- Und es gibt noch einige andere Probleme zu lösen.
  - Wir müssen eine überprüfbare Modellauswahl entsprechend der jeweiligen Anwendung treffen.
  - Wir müssen die V&V-Aktivitäten mit Methoden ausstatten, die den Anforderungen der Sicherheitsziele von IEC61508 und ISO26262 entsprechen. D.h. mit Einsatz z.B. von Assurance Cases (wir brauchen qualitative Nachweise, evtl. XAI)
  - und wir müssen es zu einem Konsens mit den quantitativen Nachweisen bringen (z.B. wie in VDR-AR-E 2842-61-X vorgeschlagen).

# Weitere Herausforderungen CRA: Secure and Responsible Design definieren

- Secure Software Development Life Cycle (source: Microsoft, SDL Process Guidance Version 5.2):




- Responsible Design Process with human well-being embedded (source: Peters et al. 2020):



# Development Roadmap TÜV AUSTRIA

TRUSTIF AI
by TÜV AUSTRIA Group & scch

**Work in progress:**

(1) Joint Risk Evaluation Safety / Security / AI Risks for Machinery & Medical Applications

(2) Mapping of Functional Safety requirements for machinery to AI model lifecycle controls

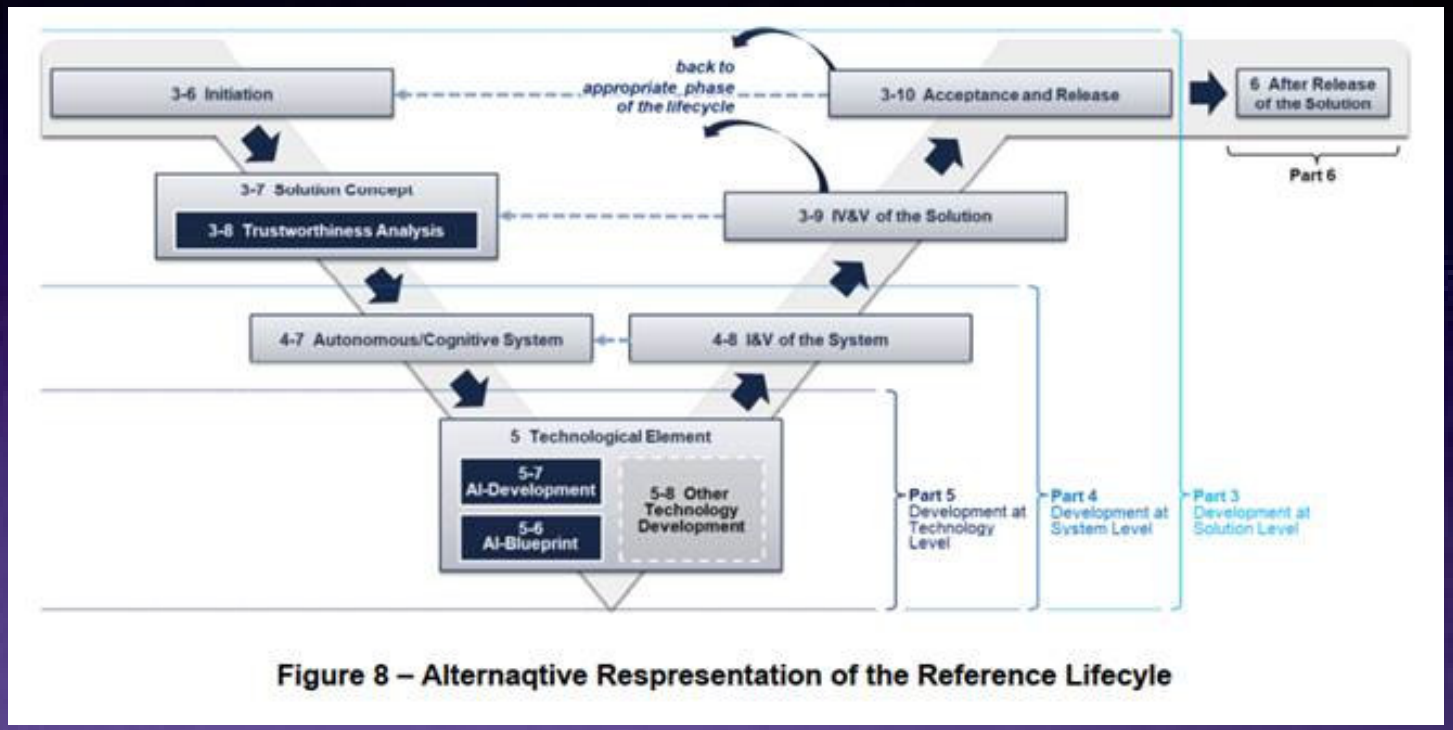
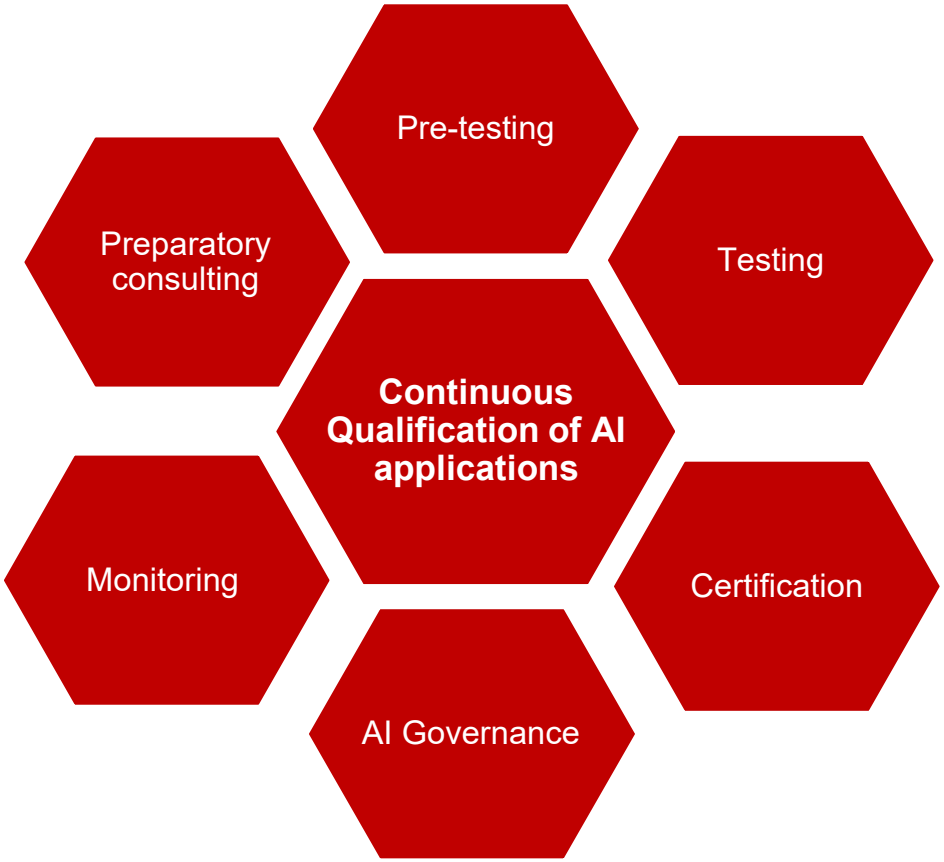



Figure 8 – Alternaqtive Respresentation of the Reference Lifecyle

# Lösungsangebot AI

TRUSTIF AI
by TÜV AUSTRIA Group & scch



## Ihr Ansprechpartner:

Thomas Doms
Geschäftsführer TRUSTIFAI GmbH
Global Product Lead AI Services TÜV AUSTRIA

E-Mail: thomas.doms@trustifai.at oder
thomas.doms@tuv.at
Mobile: +43 664 604546313
www.trustifai.at