KI-Akademie OWL: Forschungs- und Handlungsempfehlungen zur KI-Sicherheit

Hamada M. Zahera (Universität Paderborn)
Axel-Cyrille Ngonga Ngomo (Universität Paderborn)
Barbara Hammer (Universität Bielefeld)
Markus Lange-Hegermann (Technische Hochschule Ostwestfalen-Lippe)
Wolfram Schenck (Hochschule Bielefeld)
Volker Lohweg (Technische Hochschule Ostwestfalen-Lippe)

Executive Summary

Dieses Strategiepapier der KI-Akademie OWL bündelt die Forschungskompetenz von vier Hochschulen zum Thema Sicherheit künstlicher Intelligenz (KI) und empfiehlt folgende Schlüsselmaßnahmen: 1) Vernetzung regionaler Forschungseinrichtungen und Living Labs: Um technische und gesellschaftliche Herausforderungen im Bereich KI-Sicherheit und -Robustheit gemeinsam anzugehen. 2) Forschungs- und Pilotprojekte zu inklusiven und adaptiven KI-Modellen: Mit besonderem Fokus auf der Nutzung großer Sprachmodelle für Menschen mit kognitiven Beeinträchtigungen. 3) Klare und zielgruppenspezifische Handlungsempfehlungen: Für Politik, Wissenschaft und Industrie, um „Safety by Design" und „Trust by Use" bundesweit zu stärken und europäische Standards umfassend umzusetzen. Generell fasst das Papier vorhandenes Fachwissen und spezifische Schwachstellen in technischen und sozialen KI-Anwendungen zusammen. Es stellt mathematische, technologische, kommunikative und ethische Herausforderungen dar und leitet konkrete Maßnahmen ab, die für sichere, vertrauenswürdige und inklusive KI-Systeme erforderlich sind.

1 Einleitung

Sicherheit von KI umfasst *Security* (Schutz gegen Angriffe) und *Safety* (Schutz vor unbeabsichtigten Schäden). Security bezieht sich auf Maßnahmen gegen gezielte Manipulationen wie Backdoor- oder Prompt-Injection-Attacken

sowie Datenschutzverletzungen. Safety adressiert systemische Risiken wie fehlerhafte Outputs, Fairness, den Umgang mit Fehl- und Desinformation und die Bewahrung der Kontrolle durch Nutzer*innen. Die EU-Anforderungen für vertrauenswürdige KI – insbesondere der *AI Act*[1] – setzen verbindliche Standards in beiden Bereichen.[2]

Das Papier gliedert sich wie folgt: Abschnitt 2 stellt die Forschungsschwerpunkte der vier beteiligten Institutionen vor. Abschnitt 3 diskutiert domänenspezifische Schwachstellen. Abschnitt 4 fasst übergreifende technische Anforderungen zusammen. Abschnitt 5 leitet Handlungsbedarfe und Empfehlungen ab.

2 Forschungsschwerpunkte und gemeinsame Perspektiven der Partnerinstitutionen

Die vier Hochschulen der KI-Akademie OWL – die Universität Bielefeld, die Universität Paderborn, die Hochschule Bielefeld (HSBI) und die Technische Hochschule Ostwestfalen-Lippe (TH OWL) – bündeln ihre Kompetenzen in einem gemeinsamen Forschungsrahmen zur **„sicheren, vertrauenswürdigen und inklusiven KI"**. Im Folgenden werden diese gemeinsamen Achsen und methodischen Schwerpunkte beschrieben (Abb. 1).

Gemeinsame Forschungsachsen und methodische Schwerpunkte

- **Mensch–KI-Interaktion und Safety (Universität Bielefeld):** Die Forschung in Bielefeld widmet sich der *Mensch–KI-Interaktion* mit Fokus auf *Safety-by-Design*. Ziel ist es, den Schutz des Menschen vor unbeabsichtigten Risiken durch KI-Systeme sicherzustellen und vertrauensbildende, adaptive Assistenzsysteme zu entwickeln. Diese Systeme werden direkt in regionalen Bildungs- und Produktionsumgebungen in OWL eingesetzt. Forschungsfelder umfassen verhaltensadaptive Robotik, sensorbasierte Lernumgebungen und erklärbare Feedbackmechanismen.
- **Robuste und inklusive Sprachmodelle (Universität Paderborn):** Paderborn erforscht *Large Language Models* als neue kritische Infrastruktur der Wissensvermittlung. Im Mittelpunkt stehen Sicherheit, Fair-

[1] Europäisches Parlament und Rat (2024): *Verordnung (EU) 2024/1689* über harmonisierte Vorschriften für Künstliche Intelligenz (AI Act), Amtsblatt der EU L 168/1, 12.07.2024. Online verfügbar unter: `https://eur-lex.europa.eu/legal-content/DE/TXT/?uri=CELEX:32024R1689`

[2] Siehe auch OECD (2019): *OECD Principles on Artificial Intelligence*. Online verfügbar unter: `https://oecd.ai/en/ai-principles`, sowie High-Level Expert Group on Artificial Intelligence (2019): *Ethics Guidelines for Trustworthy AI*. Europäische Kommission, Brüssel. Online verfügbar unter: `https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai`

ness und Barrierefreiheit durch adaptive Antwortstrategien, Angriffsdetektion (z. B. Prompt Injection, adversariale Manipulation) und sprachliche Vereinfachung. Das Ziel ist die Entwicklung robuster, inklusiver Sprachmodelle, die in OWL insbesondere Menschen mit kognitiven Einschränkungen sowie Bildungseinrichtungen und lokale Unternehmen unterstützen.

- **Angewandte KI in technischen und sozialen Systemen (HSBI):** HSBI erforscht *Privacy-by-Design* und den Schutz personenbezogener Daten in realen sozio-technischen Kontexten wie Pflege, Bildung und Industrie 4.0. Hierbei steht die Sicherstellung von Datenschutz, sozialer Resilienz und ethischer Verantwortung beim KI-Einsatz im Vordergrund. Ziel ist es, regionale Akteure in OWL bei der sicheren und verantwortungsvollen Nutzung von KI zu befähigen.
- **Bias-Quantifizierung und Unsicherheitsmodellierung (TH OWL):** TH OWL entwickelt Methoden zur Quantifizierung von probabilistischen und possibilistischen mit dem Blick auf aleatorische und epistemische Unsicherheiten und Bias-Detektion für robuste, autonome und industrielle (Industrie 4.0) Systeme. Die Forschung trägt dazu bei, Unsicherheiten und Verzerrungen systematisch zu identifizieren und zu reduzieren, um Innovationen aus OWL verlässlich, fair, erklärbar und sicher im Kontext sozio-technischer Systeme zu gestalten.

Analytische Synthese: Interdisziplinäre Muster und Ergänzungen

Die folgenden Dimensionen verdichten die zuvor beschriebenen Forschungsschwerpunkte zu einem integrativen Rahmen, der gemeinsame Muster und ergänzende Perspektiven sichtbar macht:

(1) Vom technischen zum sozio-technischen Sicherheitsverständnis. Während Bielefeld und TH OWL primär technische und technologische Risiken adressieren, ergänzen Paderborn und HSBI dies durch gesellschaftliche und kognitive Dimensionen wie Inklusion, Nutzerkontrolle und Datenethik. Die Verbindung beider Ansätze schafft eine neue Sicherheitssemantik, die Technik und Mensch als wechselseitige Faktoren von Vertrauenswürdigkeit begreift.

(2) Safety-by-Design und Trust-by-Use. Die Hochschulen verbinden präventive (Safety-by-Design) und adaptive (Trust-by-Use) Ansätze. Technische Modelle werden nicht nur geprüft, sondern durch partizipative Verfahren mit Nutzer*innen rückgekoppelt. So entsteht ein „lebendes Sicherheitsmodell", das kontinuierlich lernt und gesellschaftlich legitimiert wird.

(3) Reallabore als transdisziplinäre Testfelder. OWL bietet mit Industrie- und Bildungspartnern die ideale Umgebung für *Living Labs*. Hier werden

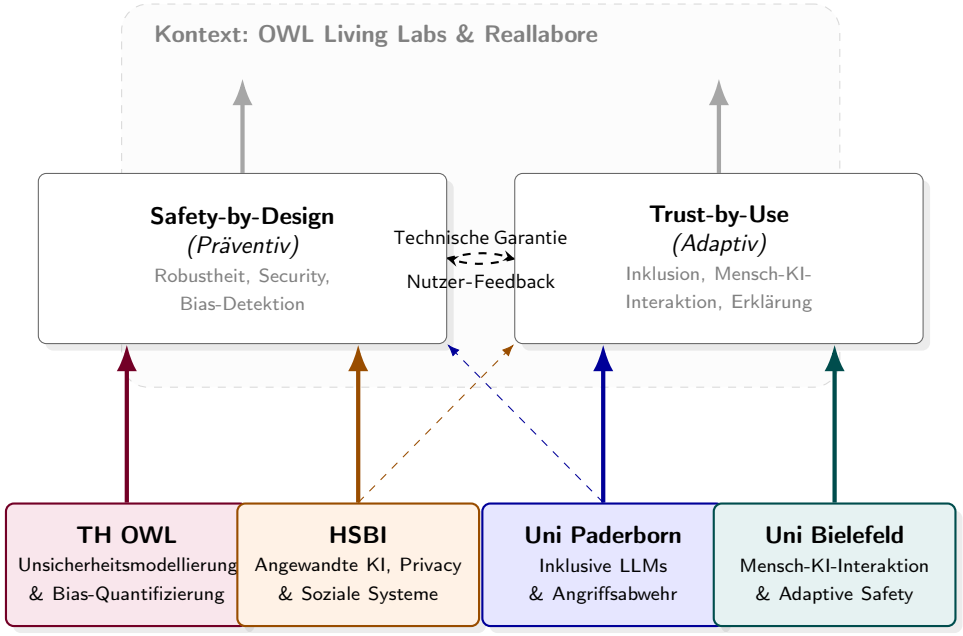

Fig. 1: Integrierter Forschungsansatz der KI-Akademie OWL: Synthese aus technischer Robustheit (Safety-by-Design) und sozialer Validierung (Trust-by-Use).

Methoden der Partnerhochschulen zusammengeführt: Bielefelds Interaktionsmodelle, Paderborns Sprach-LLMs, HSBIs Datenschutzlösungen und TH OWLs Unsicherheitsanalysen. Diese Reallabore schaffen messbare Evidenz, wie KI-Sicherheit praktisch umsetzbar ist – ein zentraler Mehrwert gegenüber rein theoretischen Ansätzen.

Gemeinsame Forschungsachsen: Praktische Anwendungsszenarien

- **Mensch–KI-Interaktion und Safety (Universität Bielefeld):** Entwicklung verhaltensadaptiver Roboter zur Unterstützung älterer Menschen in Pflegeeinrichtungen in OWL. *Szenario: Ein Living Lab testet KI-basierte Assistenzsysteme in einer lokalen Klinik zur Verbesserung der Patientensicherheit.*
- **Robuste und inklusive Sprachmodelle (Universität Paderborn):** Adaptive Large Language Models für barrierefreie digitale Lernplattformen in regionalen Bildungseinrichtungen. *Use Case: Ein Pilotprojekt für personalisierte, vereinfachte Lernbegleitung bei Weiterbildungskursen für Menschen mit kognitiven Einschränkungen.*
- **Angewandte KI in technischen und sozialen Systemen (HSBI):** Privacy-

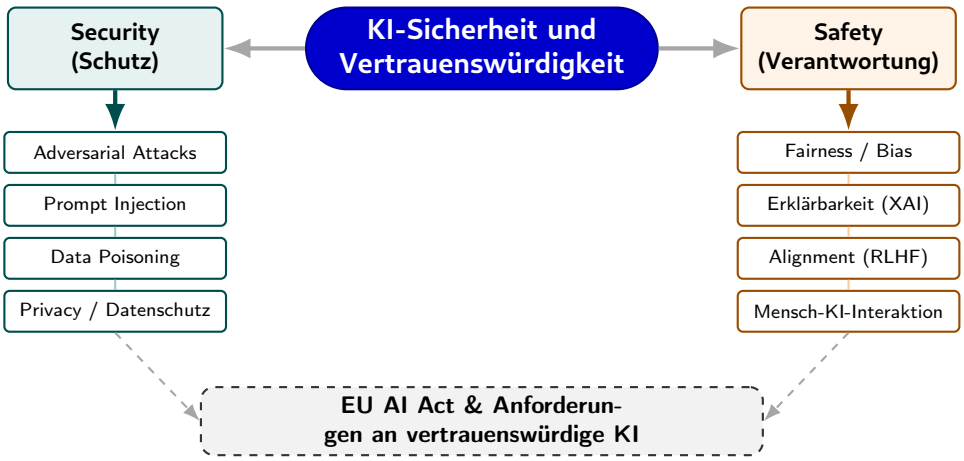

Fig. 2: Modell für Sicherheit und Vertrauenswürdigkeit: Strukturierte Übersicht im Kontext des EU AI Acts.

by-Design Lösungen für Industrie 4.0 und soziale Dienste. *Szenario: Sichere KI-basierte Datenverarbeitung beim Pflegemanagement – eine regionale Sozialagentur testet einen Prototyp zum Schutz sensibler Gesundheitsdaten.*

- **Bias-Quantifizierung und Unsicherheitsmodellierung (TH OWL):** Automatisierte Unsicherheitsmodellierung in Fertigungsanlagen und Industriesteuerungen sowie sozio-technoischen Systemenin OWL. *Use Case: Ein mittelständisches Unternehmen gestaltet die KI-Qualitätskontrolle für industrielle Robotik mit Methoden zur Unsicherheitsbewertung.*

Schlussfolgerung: Regionale Stärken als Modell für Europa

Durch die enge Verzahnung von Safety, Fairness und Inklusion entsteht ein regional verankerter, aber europaweit anschlussfähiger Forschungsrahmen, der explizit mit den Leitlinien des EU AI Acts und den OECD-Prinzipien kompatibel ist.[3] OWL positioniert sich damit als Testregion für verantwortungsvolle KI, in der wissenschaftliche Exzellenz mit gesellschaftlicher Wirkung kombiniert wird. Das gemeinsame Ziel besteht darin, **KI-Sicherheit als kulturelle und technologische Innovation zugleich** zu begreifen – ein Ansatz, der europäische Standardisierung, industrielle Transferfähigkeit und demokratische Teilhabe miteinander verbindet.

[3] Vgl. Verordnung (EU) 2024/1689, OECD (2019), HLEG (2019).

Tab. 1: Schnittstellenanalyse der Forschungsschwerpunkte: Von der Domäne zur KI-Sicherheitsdimension

Forschungsdomäne	Beispielhafte Risiken	Technische Gegenmaßnahmen	Gesellschaftliche Relevanz
Mensch–KI-Interaktion	Fehlverhalten durch missverständliche Anweisungen, Überabhängigkeit	Safety-by-Design, erklärbare KI	Vertrauen, Akzeptanz
Große Sprachmodelle	Bias, Halluzination, Prompt Injection	Red Teaming, RLHF, Datenkurierung	Barrierefreiheit, Inklusion
Industrielle Systeme	Datenmanipulation Sicherheitslücken	Privacy-by-Design, Anomaly Detection	Datenschutz, Resilienz
Gesundheitswesen	Fehlerhafte Diagnosen, Datenschutzverstöße	Bias-Korrektur, Model Monitoring	Patientensicherheit

3 Vulnerabilitäten ausgewählter Anwendungsfelder

Intelligente Tutorsysteme: Risiken durch Personalisierungsbias, Beeinträchtigung akademischer Integrität, Überabhängigkeit von KI-generierten Feedbacks.

Sozio-technische Systeme: Evasion-Attacken, die Trainingsdaten rekonstruieren oder Vorschriften umgehen, sowie Manipulation von Kommunikationswegen zwischen Mensch und KI durch Datenmissbrauch.

Große Sprachmodelle Inklusion: Anfälligkeit für Prompt Injections, Halluzinationen, Bias-Verstärkung und mangelnde Barrierefreiheit, was zu Diskriminierung führen kann.

Kognitive Automatisierung Robotik: Physische Angriffe, Fehlfunktionen und unerwartetes Verhalten mit potenziellen Sicherheitsgefahren für Mensch und Umwelt.

Medizin und Lebenswissenschaften: Angriffe durch Datenmanipulation (Data Poisoning), Bias in Diagnosemodellen, Verletzung von Datenschutz bei sensiblen Patientendaten.

Kreativwirtschaft und Medien: Deepfakes, Urheberrechtsverstöße, individualisierte und massenhafte Verbreitung von qualitativ minderwertigen bzw. extremen Inhalten.

Verwaltung: Risiken durch Datenschutzverletzungen, fehlerhafte Entscheidungsautomatisierung, juristische Haftungsfragen bei falscher Information.

Autonomes Fahren: Verlassen der Öperational Design Domain", Angriffe auf Sensorik (z.B. Spoofing), Fehlklassifikation in unbekannten Umgebungen.

Menschen mit Behinderungen: Barrieren in Nutzungsoberflächen, diskriminierender Bias, mangelnde Individualisierung bei KI-Modellen.

IT-Sicherheit: Großskalige Impersonation-Angriffe mittels Deepfakes und Automation von Cyberangriffen durch generative KI-Modelle.

4 Technische Anforderungen und Lösungsansätze

4.1 Übergreifende technische Anforderungen

- **Fairness/Bias**: Erkennung und Mitigation von Daten- und Modellbias (z.B. Bias-Detektion mittels Equalized Odds, Adversarial Debiasing).
- **Operationalisierung von Limitationen**: Dokumentation und testbare Definition zulässiger Systemgrenzen (Operational Design Domain, Model Cards).
- **Umgang mit Fehlern**: Mechanismen für Fehlervorhersage (Kalibrierung, Modellierungsansätze), Graceful Degradation und effizientes Error-Handling.
- **Alignment mit menschlichen Werten**: RLHF, Supervised Fine-Tuning, Red Teaming sowie Herausforderungen bei deren Umsetzung.

4.2 Beispiel: Inklusive Sprachmodelle (Universität Paderborn)

Die Universität Paderborn entwickelt Verfahren zur Detektion und automatischen Abwehr unsicherer oder manipulierter Eingaben (Prompt Injection, adversariale Angriffe) und setzt dabei auf adaptive Antwortstrategien (Antwortverweigerung, Rückfragen, Reformulierung). Sprachvereinfachung, kontinuierliche Kalibrierung und systematische Fairnessprüfungen maximieren die Nutzungsfreundlichkeit und Robustheit.

5 Handlungsbedarfe und Empfehlungen

Die Analyse der technischen und gesellschaftlichen Herausforderungen zeigt klaren Handlungsbedarf auf mehreren Ebenen. Die folgenden Empfehlungen (R1–R6) adressieren Akteure aus Wissenschaft, Unternehmen und Politik spezifizieren jeweils: *wer handeln sollte*, *was zu tun ist* und den *erwarteten Nutzen.*

1. **R1 – Einheitliche Definition von KI-Sicherheitszielen**

   Wer sollte handeln: Landesregierung NRW, KI-Akademie OWL und beteiligte Hochschulen in Kooperation mit Normungsorganisationen (DIN, ISO).

   Was ist zu tun: Erarbeitung eines einheitlichen Referenzrahmens für *Safety*, *Security* und *Trustworthiness* im Einklang mit dem EU AI Act[4]. Dieser Rahmen soll regionale Standards und Zertifizierungskriterien für sichere KI-Anwendungen festlegen.

   Erwarteter Nutzen: Fördert die Vergleichbarkeit von Projekten, erhöht regulatorische Sicherheit für Unternehmen und stärkt OWL als Modellregion für vertrauenswürdige KI.

2. **R2 – Integration von Safety-by-Design in Forschungsförderung**

   Wer sollte handeln: BMFTR, MWIDE NRW, Projektträger Jülich, Forschungseinrichtungen.

   Was ist zu tun: Förderprogramme sollen bereits in der Konzeptphase Anforderungen an Risikobewertung, *Alignment* und Fairness vorsehen. Evaluierungskriterien in Ausschreibungen sind entsprechend anzupassen.

   Erwarteter Nutzen: Sichert frühzeitige Integration ethischer und technischer Sicherheitsaspekte und reduziert Entwicklungsrisiken in geförderten Projekten.

3. **R3 – Ausbau von Test- und Zertifizierungsinfrastrukturen**

   Wer sollte handeln: Zertifizierungsstellen, Landescluster IT.NRW, Hochschulen und Industriepartner.

   Was ist zu tun: Aufbau regionaler Testbeds und Reallabore in OWL für sichere, barrierefreie und robuste KI-Systeme. Integration in bestehende Prüfinfrastrukturen (z. B. Labs von HSBI und TH OWL).

   Erwarteter Nutzen: Ermöglicht messbare Evidenz für KI-Sicherheit, unterstützt die praktische Umsetzung des EU AI Acts und stärkt regionale Innovationsökosysteme.

[4] Vgl. Verordnung (EU) 2024/1689.

4. **R4 – Verknüpfung von technischer und ethischer Ausbildung**

   Wer sollte handeln: Hochschulen, Kultusministerium NRW, Industriepartner (Dual-Studiengänge).

   Was ist zu tun: Integration von Ethik, Recht und gesellschaftlicher Verantwortung in KI-Studiengänge, Weiterbildung und Ingenieursausbildung. Kooperative Lehrmodule zwischen Informatik, Sozialwissenschaften und Recht aufbauen.

   Erwarteter Nutzen: Schafft generationenübergreifend Kompetenzen für verantwortungsvolle KI-Entwicklung und erhöht gesellschaftliche Akzeptanz.

5. **R5 – Aufbau einer Daten-Governance für Inklusion**

   Wer sollte handeln: Landesdatenschutzbeauftragte, Hochschulen, kommunale Verwaltungen und Forschungsnetzwerke.

   Was ist zu tun: Entwicklung von Leitlinien und Metadatenstandards für diversitätsgerechte Datensätze, die alle gesellschaftlichen Gruppen angemessen repräsentieren. Einrichtung eines zentralen Datenkompetenzzentrums OWL.

   Erwarteter Nutzen: Reduziert Verzerrungen in Trainingsdaten, verbessert Fairness und Barrierefreiheit von KI-Systemen und unterstützt Inklusionsziele nach EU-Richtlinie.

6. **R6 – Transferplattform für vertrauenswürdige KI**

   Wer sollte handeln: KI-Akademie OWL, Wirtschaftsförderung NRW, kommunale Verwaltungen, Industriecluster.

   Was ist zu tun: Einrichtung einer offenen digitalen Plattform zur Vernetzung von Forschung, Industrie und öffentlicher Verwaltung. Bereitstellung von Best-Practice-Beispielen, Schulungsmaterialien und Testfällen zur Umsetzung der KI-Sicherheitsprinzipien.

   Erwarteter Nutzen: Fördert den Wissenstransfer, stärkt die Innovationsfähigkeit regionaler KMU und beschleunigt den Einsatz sicherer und vertrauenswürdiger KI-Lösungen.

Diese Maßnahmen unterstützen die Umsetzung internationaler Leitlinien – des EU AI Acts, der OECD-Prinzipien und der HLEG-Leitlinien für vertrauenswürdige KI – und tragen dazu bei, OWL als Modellregion für verantwortungsvolle KI zu etablieren.

Glossar

Prompt Injection: Spezialform der Manipulationseingabe, mit der Sicherheitsmechanismen oder Nutzungsbeschränkungen von Sprachmodellen um-

gangen werden.

Red Teaming: Systematisches Testen von KI-Systemen durch simulierte Angriffe, um Schwachstellen, Fehlverhalten und ethische Risiken zu identifizieren.

Fairness/Bias: Zustand, in dem KI-Modelle keine diskriminierenden Verzerrungen gegenüber geschützten Gruppen aufweisen. Bias kann durch unausgewogene Trainingsdaten, Modellarchitekturen oder Nutzungskontexte entstehen.

Safety-by-Design: Gestaltungsprinzip, bei dem Sicherheits-, Ethik- und Vertrauensaspekte bereits in frühen Entwicklungsphasen von KI-Systemen berücksichtigt werden.

Trust-by-Use: Ansatz, nach dem Vertrauen in KI-Systeme durch praktische Erfahrung, Transparenz und kontinuierliche Evaluation im Einsatzkontext entsteht.

Living Lab: Reallabor, das Forschung, Industrie und Gesellschaft in einem realen Anwendungskontext zusammenführt, um KI-Lösungen gemeinsam zu erproben und zu bewerten.

Operational Design Domain (ODD): Definierter Rahmen, in dem ein KI-System sicher betrieben werden kann; umfasst Umgebungsbedingungen, Aufgabenbereich und technische Grenzen.

Reinforcement Learning from Human Feedback (RLHF): Trainingsverfahren, bei dem menschliche Rückmeldungen genutzt werden, um das Verhalten von Sprach- oder Entscheidungssystemen an menschliche Werte anzupassen.

Alignment: Abstimmung der Ziele und Entscheidungen eines KI-Systems auf menschliche Werte, ethische Prinzipien und gesellschaftliche Normen.

Data Governance: Gesamtheit von Richtlinien, Verantwortlichkeiten und Verfahren zur sicheren, fairen und rechtskonformen Nutzung von Daten, insbesondere für KI-Training und Evaluation.