

Whitepaper

Trustworthy AI Operations: Prinzipien, Praktiken und Herausforderungen



Whitepaper

Trustworthy AI Operations: Prinzipien, Praktiken und Herausforderungen

Autorinnen und Autoren

Lisa Fink, Fraunhofer IAIS, KI.NRW

Lennard Helmer, Fraunhofer IAIS

Fabian Malms, Fraunhofer IAIS

Claudio Martens, Fraunhofer IAIS

Michael Mock, Fraunhofer IAIS

Benny Jörg Stein, Fraunhofer IAIS

Sermad Abbas, BITMARCK

Februar 2026

www.iais.fraunhofer.de/zertifizierte-ki



KI.NRW ist die zentrale Anlaufstelle für Künstliche Intelligenz in Nordrhein-Westfalen. Die Kompetenzplattform baut das Land zu einem bundesweit führenden Standort für angewandte KI aus. Ziel ist es, den Transfer von KI aus der Spitzenforschung in die Wirtschaft zu beschleunigen und Impulse im gesellschaftlichen Dialog zu setzen. Dabei stellt KI.NRW die Menschen und ihre ethischen Grundsätze in den Mittelpunkt der Gestaltung von KI.

www.ki.nrw



Das Projekt ZERTIFIZIERTE KI fördert die Entwicklung und Standardisierung von Prüfkriterien, -methoden und -werkzeuge für KI-Systeme, um die technische Zuverlässigkeit und einen verantwortungsvollen Umgang mit der Technologie zu gewährleisten.

www.zertifizierte-ki.de

Das Fraunhofer IAIS

Das Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS mit Sitz in Sankt Augustin bei Bonn sowie Standorten in Dresden und Heilbronn zählt zu den führenden Institutionen für angewandte Forschung auf den Gebieten Künstliche Intelligenz (KI), Maschinelles Lernen und Generative KI in Deutschland und Europa.

Mit rund 350 Mitarbeitenden entwickelt das Fraunhofer IAIS Strategien, Technologien und Lösungen für Unternehmen, Behörden und Organisationen entlang der gesamten Wertschöpfungskette. Die Teams des Instituts sind auf die Anforderungen der Branchen Finanzen & Recht, Gesundheit, Handel, Industrie & Automotive, Medien sowie öffentlicher Sektor spezialisiert. Branchenübergreifend reicht das Angebot von digitaler Transformation, Generativen KI-Modellen und intelligenter Prozessautomatisierung über KI-Prüfung und -Absicherung bis hin zu Nachhaltigkeits- und Resilienzstrategien. Mit einem umfassenden Weiterbildungsprogramm, bestehend aus fundierten Data-Science- und KI-Schulungen mit Zertifizierungsoption, qualifiziert das Fraunhofer IAIS Menschen unterschiedlicher Fachrichtungen und Kenntnisstufen.

www.iais.fraunhofer.de

Inhalt

| | |
|---|-----------|
| Vorwort | 5 |
| 1 Einführung | 6 |
| 1.1 Die Prinzipien vertrauenswürdiger KI | 6 |
| 1.2 Regulatorische Herausforderungen von vertrauenswürdiger KI | 7 |
| 1.3 Unternehmensanforderungen an vertrauenswürdige KI | 8 |
| 1.4 Herausforderungen bei der Operationalisierung von vertrauenswürdiger KI | 9 |
| 1.5 Zusammenfassung | 9 |
| 2 Trustworthy AI Operations | 11 |
| 2.1 Machine Learning Operations (MLOps) | 11 |
| 2.2 Operationalisierung auf drei Ebenen | 13 |
| 2.3 Einbettung von Qualitäts- und Risikomanagement in die Entwicklung | 14 |
| 2.4 Zusammenfassung | 15 |
| 3 TAIOps-Methodik | 16 |
| 3.1 KI-Kodex und -Policy | 16 |
| 3.2 KI-Steckbrief | 16 |
| 3.3 Systembeschreibung und Architekturdokumentation | 18 |
| 3.4 Datasheet | 18 |
| 3.5 Decision Records | 19 |
| 3.6 Metrik-Katalog | 22 |
| 3.7 Zusammenfassung | 23 |
| 4 BITMARCK – KI-Governance für eine vertrauenswürdige KI | 24 |
| 4.1 Das Lebensphasenmodell | 24 |
| 4.2 Umsetzung von TAIOps bei BITMARCK | 25 |
| 4.3 Einführung des TAIOps-Leitfadens in den Projekten | 27 |
| 4.4 Erkenntnisse und Empfehlungen | 27 |
| 5 Diskussion und Ausblick | 29 |
| Autorinnen und Autoren | 30 |
| Literaturverzeichnis | 31 |
| Impressum | 32 |

Abbildungs- und Tabellenverzeichnis

| | |
|---|----|
| Abbildung 1: CRISP-DM-Vorgehen. | 11 |
| Abbildung 2: Der MLOps-Zyklus. | 12 |
| Abbildung 3: Vertrauenswürdige KI »by design«. | 13 |
| Abbildung 4: Skizze zu unternehmensweiten und projektspezifischen Elementen der TAIOps-Ebenen. | 14 |
| Abbildung 5: Hierarchiepyramide von KI-Kodex, KI-Policy und Handlungsempfehlungen. | 16 |
| Abbildung 6: Verwendung von Decision Records bei BITMARCK (Kapitel 4). | 20 |
| Abbildung 7: Zuordnung der MLOps-Phasen zu den Lebensphasen im BITMARCK KI-Lebensphasenmodell.. | 25 |
| | |
| Tabelle 1: Übersicht zu den Dimensionen eines KI-Steckbriefes im Sinne des Fraunhofer IAIS Prüfkatalogs. | 17 |
| Tabelle 2: Beispiele für einige der Fragen..... | 19 |
| Tabelle 3: Fiktives Beispiel für einen Decision Record..... | 21 |

Vorwort

Spätestens mit dem Erfolg von KI-Systemen, wie beispielsweise ChatGPT von OpenAI, überlegen viele Unternehmen eigene KI-Systeme zu entwickeln und ihren Mitarbeitenden oder Kundinnen und Kunden zur Verfügung zu stellen, wie aktuelle Erhebungen des Branchenverbandes Bitkom zeigen¹ – beispielsweise, um von potenziellen Effizienzgewinnen zu profitieren oder sich mit modernen Produkten am Markt zu platzieren. Ein etabliertes Vorgehen zur Entwicklung von KI-Systemen ist »Machine Learning Operations« (MLOps), welches sich an bewährten Methoden aus der Softwareentwicklung orientiert und besonderen Wert auf Qualität und Effizienz legt.

Mit der zunehmenden Integration von KI in den beruflichen und privaten Alltag gewinnen Fragestellungen zur Vertrauenswürdigkeit von KI-Systemen zunehmend an Bedeutung. Unternehmen sehen sich mit Anforderungen im Hinblick auf Fairness, Transparenz und Datenschutz konfrontiert. Mit der Verabschiedung des EU AI Acts hat auch der europäische Gesetzgeber reagiert und verlangt den Nachweis, dass KI-Systeme die Gesundheit, Sicherheit und Grundrechte ihrer Nutzer gewährleisten, insbesondere wenn sie in Hochrisikobereichen eingesetzt werden sollen.

Ziel dieses Whitepapers ist es, die Anforderungen an die Entwicklung vertrauenswürdiger KI-Systeme zu analysieren und darzustellen, wie diese durch die Integration geeigneter Verfahren und Methodiken in das MLOps-Vorgehen integriert werden können. Das daraus abgeleitete, erweiterte Vorgehen bezeichnen wir als Trustworthy AI Operations (TAIOps).

Dieses Whitepaper wurde im Rahmen des KI.NRW-Flagshipprojektes »ZERTIFIZIERTE KI« gefördert.

¹ Bitkom e.V., 2025.

1 Einführung

Aktuelle Erhebungen des Branchenverbands Bitkom² zeigen, dass sich zwar mehr als die Hälfte der Unternehmen in Deutschland mit KI beschäftigen, jedoch nur etwa 17–20 Prozent diese auch produktiv einsetzen. 64 Prozent der Unternehmen sehen sich selbst als Nachzügler, 22 Prozent als bereits abgehängt. Gleichzeitig betrachten 78 Prozent der Unternehmen KI als Chance und Dreiviertel der Bevölkerung befürworten ihren Einsatz, beispielsweise im Gesundheitswesen.

Die hohe Diskrepanz zwischen der grundsätzlichen Auseinandersetzung mit KI und dem Gefühl bereits abgehängt worden zu sein zeigt, dass KI-Einsatz und -Entwicklung eine Herausforderung darstellen. Die dazu notwendigen Praktiken und Methodiken sind unbekannt oder es fehlt die Expertise, diese produktiv einsetzen zu können. Während beim reinen Einsatz von KI potenziell auf leistungsstarke Serviceanbieter zurückgegriffen werden kann, ist die unternehmensinterne KI-Entwicklung komplizierter. Die hohe Attraktivität, insbesondere durch die Hoheit über die eigenen Daten und die Vermeidung von Abhängigkeiten von externen Unternehmen, wird begleitet von hohem Bedarf an Experten, Fallstricken im Training und Betrieb von KI-Modellen und Aufwänden zum Aufbau einer geeigneten Infrastruktur.

Ein in der Praxis etabliertes Vorgehensmodell zur Entwicklung von KI-Systemen ist MLOps. Es basiert auf einem Paradigma, das die systematische Entwicklung, Integration und Überwachung von KI-Systemen durch strukturierte Methoden, Werkzeuge und Prozesse ermöglicht. Der Fokus liegt dabei auf der Effizienz der Entwicklungsprozesse sowie der Qualität der resultierenden Softwareartefakte. Aspekte wie Vertrauenswürdigkeit und Gesetzeskonformität sind in den aktuellen Ausprägungen von MLOps jedoch unzureichend berücksichtigt.

Dieses Whitepaper zeigt Methoden auf, die diese Lücke adressieren und unter dem Begriff TAIOps (»Trustworthy AI Operations«) gesammelt werden. Ziel ist es, KI-Prozesse nicht nur effizient und skalierbar, sondern auch gesetzeskonform, sicher, transparent und ethisch verantwortungsvoll zu gestalten. Dies ist inhaltlich, technisch und organisatorisch anspruchsvoll und selbst erfahrene Entwicklungsteams benötigen hierfür Hilfestellungen.

Die Operationalisierung dieser Zielsetzung erfolgt entlang zentraler Dimensionen der Vertrauenswürdigkeit, welche spezifische Anforderungen an die Gestaltung und Entwicklung von KI-Systemen stellen.

Dabei werden Fragen beantwortet, wie beispielsweise:

- Wie kann das System beschrieben und dokumentiert werden? (Kapitel 3.2)
- Wann müssen Entscheidungen getroffen und protokolliert werden? (Kapitel 3.5)
- Welche Datenströme sind relevant und müssen dokumentiert werden? (Kapitel 3.4)
- Wie wird die Design- und Architekturdokumentation strukturiert? (Kapitel 3.3)

Darüber hinaus ist die Qualität und Vertrauenswürdigkeit von KI-Systemen in hohem Maße von dem KI-Anwendungsfall und dessen Kontext abhängig, wie beispielsweise den verwendeten Daten für das Modelltraining und -testen.

Mit der Verabschiedung des EU AI Acts im August 2024 wurde ein einheitlicher Rechtsrahmen für die Entwicklung, Bereitstellung und Nutzung von Künstlicher Intelligenz in Europa geschaffen. Die Verordnung enthält einen umfassenden Anforderungskatalog, insbesondere für Anbieter von Hochrisiko-KI-Systemen, und betont den Schutz der Anwendenden. Entwickler und Anbieter sind verpflichtet, die Vertrauenswürdigkeit ihrer Systeme über den gesamten Lebenszyklus hinweg sicherzustellen und zu dokumentieren.

Diese gesellschaftliche und regulatorische Entwicklung hat die Erwartungshaltung in der Industrie an KI-Systeme verändert und erhöht den Bedarf an Methodiken zur Sicherstellung von Vertrauenswürdigkeit und deren Einbettung in die existierenden Entwicklungsprozesse, wie MLOps. Vor diesem Hintergrund wurde das vorliegende Whitepaper erarbeitet, mit der Motivation geeignete Methodiken aufzuzeigen und zu besprechen. Zusätzlich wird anhand eines konkreten Use Cases mit dem Unternehmen BITMARCK verdeutlicht, wie diese Methodiken in Unternehmen in existierende Entwicklungsprozesse integriert werden können und auf mögliche Herausforderungen hingewiesen.

1.1 Die Prinzipien vertrauenswürdiger KI

Bevor konkrete Umsetzungsmöglichkeiten diskutiert werden können, ist eine präzise Definition des Begriffs »Vertrauenswürdige Künstliche Intelligenz (KI)« erforderlich.

² Streim & Thomas, 2025.

Die Europäische Union hat durch die hochrangige Experten­gruppe für Künstliche Intelligenz grundlegende Anforderungen formuliert, die erfüllt sein müssen, damit ein KI-System als vertrauenswürdig gilt. Auch in der wissenschaftlichen Forschung hat sich ein dynamisches Feld etabliert, das sich mit der systematischen Untersuchung der Vertrauenswürdigkeit von KI-Systemen befasst. Zentrale Fragestellungen sind unter anderem:

- Wie lässt sich die Fairness eines KI-Systems überprüfen?
- Wie kann die zweckgemäße Funktion eines KI-Systems sichergestellt werden?
- Welche Maßnahmen schützen KI-Systeme vor Fehlverhalten oder Manipulation?
- Wie können Entscheidungen von KI-Systemen nachvollziehbar gemacht werden?
- Wie lässt sich der zugrunde liegende Datenfluss transparent darstellen und dokumentieren?

Die Beantwortung dieser Fragen ist kontextabhängig und erfordert eine differenzierte Betrachtung des jeweiligen Einsatzszenarios. In der Folge wurden zahlreiche Methoden und Metriken entwickelt, die einzelne Aspekte der Vertrauenswürdigkeit adressieren.

Eine umfassende Analyse dieser Aspekte wurde von den Forschenden des Fraunhofer-Instituts für Intelligente Analyse- und Informationssysteme IAIS durchgeführt. Im Rahmen dieser Studie wurde ein Prüfkatalog entwickelt, der sechs Dimensionen der Vertrauenswürdigkeit sowie über 200 Prüfkriterien definiert.³ Diese Dimensionen bilden thematische Kategorien, die für die Bewertung von KI-Systemen essenziell sind:

Datenschutz

KI-Systeme sind in hohem Maße datengetrieben. Die Verarbeitung personenbezogener und geschäftsrelevanter Daten erfordert strenge Maßnahmen zum Schutz sensibler Informationen im Einklang mit den Regelungen der Datenschutzgrundverordnung und anderer geltender Regelungen. Datenschutz ist daher ein zentrales Element vertrauenswürdiger KI, insbesondere im Hinblick auf die Trainings-, Test- und Validierungsdaten, sowie Ein- und Ausgaben im operativen Betrieb.

Sicherheit

Diese Dimension umfasst technische und organisatorische Maßnahmen zur Absicherung von KI-Systemen gegen unbefugten Zugriff, Manipulation und Systemversagen. Sie berücksichtigt sowohl die funktionale Sicherheit als auch den Schutz physischer Umgebungen, sofern das KI-System darauf Einfluss nimmt.

Robustheit

Ein KI-System gilt als robust, wenn es zuverlässig innerhalb definierter Anwendungsgrenzen agiert und angemessen auf unbekannte Eingaben oder Ausreißer reagiert. Die Robustheit umfasst auch die Bewertung des Systemverhaltens außerhalb der vorgesehenen Einsatzbereiche.

Transparenz

Die Nachvollziehbarkeit von Entscheidungen ist insbesondere bei automatisierten Prozessen von zentraler Bedeutung. Transparenz ermöglicht sowohl die Kontrolle der Modellausgaben als auch die rechtssichere Dokumentation der Entscheidungsprozesse.

Fairness

Da KI-Systeme auf historischen Daten basieren, besteht die Gefahr der Reproduktion diskriminierender Muster. Die Dimension Fairness befasst sich mit der Identifikation und Bewertung solcher Verzerrungen in Daten und Modellausgaben.

Autonomie & Kontrolle

Die Interaktion zwischen KI-Systemen und menschlichen Entscheidungsträgern muss so gestaltet sein, dass die finale Entscheidungsgewalt stets beim Menschen verbleibt. Menschliche Akteure müssen in der Lage sein, KI-Entscheidungen zu hinterfragen und gegebenenfalls zu revidieren.

Diese sechs Dimensionen ermöglichen eine ganzheitliche Betrachtung der Vertrauenswürdigkeit von KI-Systemen und bilden die konzeptionelle Grundlage für das Verständnis von vertrauenswürdiger KI. Ein KI-System kann nur dann als vertrauenswürdig betrachtet werden, wenn es in all diesen Bereichen geeignete Maßnahmen zur Risikominderung implementiert. Unternehmen müssen sorgfältig abwägen, welches Risikoniveau sie beim Einsatz von KI-Systemen als akzeptabel erachten und ihr Risikomanagement entsprechend ausrichten.

1.2 Regulatorische Herausforderungen von vertrauenswürdiger KI

Die zunehmende Verbreitung von KI-Systemen in gesellschaftlich relevanten Bereichen hat die Europäische Union dazu veranlasst, mit dem EU AI Act eine umfassende Regulierung für KI zu schaffen. Ziel dieser Verordnung ist die Förderung einer menschenzentrierten und vertrauenswürdigen KI, die sowohl rechtlichen als auch ethischen Anforderungen gerecht wird. Bereits im Vorfeld des Gesetzgebungsverfahrens wurde die »High-Level Expert Group on Artificial Intelligence« (AI HLEG) eingesetzt. Sie entwickelte Empfehlungen und ethische Leitlinien für die Entwicklung vertrauenswürdiger KI-Systeme und

³ Poretschkin, et al. 2021.

konkretisierte den Begriff der »Vertrauenswürdigkeit« durch drei zentrale Anforderungen, die über den gesamten Lebenszyklus eines KI-Systems hinweg erfüllt sein müssen:

1. Rechtmäßigkeit

KI-Systeme müssen mit geltendem Recht und regulatorischen Vorgaben in Einklang stehen.

2. Ethik

Die Entwicklung und Anwendung von KI müssen ethischen Grundsätzen folgen, insbesondere der Achtung der Grund- und Menschenrechte.

3. Robustheit

KI-Systeme müssen technisch und sozial zuverlässig sein und dürfen keine unbeabsichtigten Schäden verursachen.

Diese Anforderungen wurden durch die Differenzierung zwischen ethischen Grundsätzen, wie menschlicher Autonomie, Schadensvermeidung und Fairness weiter konkretisiert. Das daraus abgeleitete Konzept der Vertrauenswürdigkeit wurde als umfassendes Qualitätsmerkmal in den EU AI Act integriert. Dieser verfolgt einen risikobasierten Ansatz, bei dem KI-Systeme in verschiedene Risikokategorien eingeteilt werden. Nur bestimmte KI-Anwendungen (z. B. Emotionserkennung am Arbeitsplatz) sind verboten. Für bestimmte Bereiche, die ein Risiko bergen (z. B. im Personalwesen oder bei der Kreditvergabe), gelten strenge Regeln.

Für diese sogenannten Hochrisiko-KI-Systeme müssen daher bestimmte Anforderungen etwa hinsichtlich Genauigkeit, Robustheit und Datenqualität berücksichtigt werden.

Zur Sicherstellung der regulatorischen Anforderungen sind Anbieter von KI-Systemen mit hohem Risiko dazu verpflichtet, ein Qualitätsmanagementsystem (QMS) zu implementieren. Dieses dient der strukturierten Dokumentation und Bewertung von Designentscheidungen, Risiken und Qualitätsmerkmalen über den gesamten Lebenszyklus eines KI-Systems hinweg.

1.3 Unternehmensanforderungen an vertrauenswürdige KI

Neben extrinsischen Motivationsfaktoren, wie gesetzlichen Regulierungen und dem Erwartungsdruck externer Stakeholder, lohnt sich ein Blick auf die intrinsischen Beweggründe, die Unternehmen dazu veranlassen, sich mit der Entwicklung und dem Betrieb vertrauenswürdiger KI-Systeme auseinanderzusetzen. Die Prinzipien vertrauenswürdiger KI basieren häufig auf etablierten Best Practices der KI-Systementwicklung und

sollten idealerweise auch ohne externen Druck berücksichtigt werden. Risiken für die Vertrauenswürdigkeit – etwa in Bezug auf Vertraulichkeit, Verlässlichkeit oder Sicherheit – frühzeitig zu erkennen und zu minimieren, schützt Unternehmen vor finanziellen Verlusten und Reputationsschäden.

Das allgemeine Potenzial von KI bringt bereits zahlreiche intrinsische Motivationsfaktoren mit sich, beispielsweise mögliche Effizienzsteigerung oder die Automatisierung von repetitiven Tätigkeiten. Viele Mitarbeitende zeigen daher eine grundsätzliche Offenheit gegenüber KI-Technologien. In einer Studie⁴ berichten über 80 Prozent der Studienteilnehmenden, dass sie mehr über KI lernen möchten. Ein verantwortungsvoller Umgang mit KI sowie der gezielte Abbau von Unsicherheiten im Umgang mit diesen Systemen können als Katalysator für eine vertiefte Auseinandersetzung mit den nachfolgenden Faktoren dienen.

Risikominimierung

Die Erhöhung der Vertrauenswürdigkeit eines Systems bedeutet, potenzielle Risiken systematisch zu identifizieren und geeignete Maßnahmen zu deren Minimierung oder Eliminierung zu ergreifen. Die Operationalisierung vertrauenswürdiger KI stellt somit ein effektives Instrument des Risikomanagements dar. Unternehmen können dadurch Schäden – etwa durch fehlerhafte Entscheidungen, Systemausfälle oder Reputationsverluste – proaktiv vermeiden.

Auditierfähigkeit

In stark regulierten Märkten, wie dem Gesundheitssektor oder bei Betreibern kritischer Infrastrukturen wird erwartet, dass KI-Systeme regelmäßig auditiert werden. Der EU AI Act fordert für Hochrisikosysteme explizit eine Bestätigung der Konformität durch Audits. Systeme mit hoher Auditierfähigkeit – also solche, die den Prüfprozess bestehen und aktiv unterstützen – ermöglichen eine effizientere Durchführung von Audits und reduzieren den Aufwand für nachträgliche Dokumentationen und Korrekturen.

Standardisierung

Die Etablierung von Standards bei der KI-Entwicklung erleichtert das Onboarding neuer Entwicklerinnen und Entwickler und fördert die Wiederverwendung bewährter Lösungen. Unternehmen profitieren von effizienteren Entwicklungsprozessen und einer gesteigerten Performance.

Qualitätssicherung

Maßnahmen zur Steigerung der Vertrauenswürdigkeit von KI-Systemen basieren häufig auf anerkannten Best-Practices aus der Industrie. Ihre systematische Integration in Entwicklungsprozesse erhöht die Qualität der resultierenden Systeme und stellt eine zentrale Maßnahme zur Qualitätssicherung dar.

⁴ Gillespie, et al. 2025.

Know-how-Aufbau

Während Kompetenzen zur Entwicklung leistungsfähiger KI-Systeme meist vorhanden sind, fehlt häufig das Wissen über Methoden aus dem Bereich der vertrauenswürdigen KI. Ein strukturierter Prozess zur Risikoerkennung und -bewältigung fördert den Aufbau dieses Know-hows und ermöglicht dessen nachhaltige Verankerung im Unternehmen.

Innovationspotenzial

Durch die frühzeitige Identifikation und Mitigation von Risiken können KI-Anwendungen auch in sensiblen oder unternehmenskritischen Bereichen sicher geplant und umgesetzt werden. Dies stärkt das Vertrauen in die Technologie und erweitert den strategischen Handlungsspielraum und fördert die Nutzung von Automatisierungspotenzialen.

Verkaufsförderung

Die öffentliche Sensibilisierung für Risiken von KI-Systemen nimmt zu. Sowohl abstrakte Risiken wie eine Störung des gesellschaftlichen Friedens durch unfair agierende Systeme, als auch konkrete Risiken, wie die Verletzung des Datenschutzes oder intransparente Fehlentscheidungen bei der Verwendung, werden vermehrt in der Öffentlichkeit wahrgenommen. Unternehmen, die sich klar zur Vertrauenswürdigkeit ihrer KI-Systeme bekennen, können sich positiv von Wettbewerbern abheben.

Diese Liste ist nicht abschließend. Je nach Branche und strategischer Positionierung können weitere intrinsische Faktoren relevant sein. Eine unternehmensspezifische Analyse der Beweggründe zur Beschäftigung mit vertrauenswürdiger KI ist ratsam. Sie unterstützt die interne Kommunikation und erleichtert die Argumentation für notwendige Investitionen.

1.4 Herausforderungen bei der Operationalisierung von vertrauenswürdiger KI

Die konkrete Umsetzung der Prinzipien vertrauenswürdiger KI stellt Unternehmen vor erhebliche Herausforderungen. Die bloße Festlegung ethischer Leitlinien reicht nicht aus, um die Entwicklung und den Einsatz vertrauenswürdiger KI-Systeme wirksam zu fördern.⁵ Vielmehr bedarf es einer systematischen Übersetzung abstrakter ethischer Zielsetzungen in konkrete technische und organisatorische Anforderungen, die als Orientierung für Entwicklungsteams dienen können.

Ohne diese Operationalisierung besteht die Gefahr eines sogenannten »Greenwashing« – also der Diskrepanz zwischen öffentlich kommunizierten Zielen und tatsächlicher Umsetzung.⁶ Dies kann nicht nur das Vertrauen in KI-Systeme

untergraben, sondern auch regulatorische und reputative Risiken für Unternehmen nach sich ziehen.

Ein zentraler Schritt zur Operationalisierung besteht in der präzisen Definition des Anwendungskontexts, in dem ein KI-System eingesetzt werden soll. Wichtige Einflussfaktoren sind die rechtlichen Rahmenbedingungen, die Marktsituation, Compliance-Anforderungen und Stakeholder-Erwartungen. In diesem Zusammenhang bietet die Norm ISO 42001 einen strukturierten Rahmen für das Management von KI-Systemen. Sie unterstützt Organisationen dabei, den KI-Einsatz strategisch auszurichten und mit den übergeordneten Unternehmenszielen zu harmonisieren.

Die Norm fordert unter anderem:

- eine klare strategische Zieldefinition
- die Einbindung relevanter Stakeholder
- die Einhaltung ethischer und rechtlicher Standards
- Transparenz zur Förderung von Vertrauen

Aus dem definierten Kontext lassen sich übergeordnete Governance-Strukturen ableiten, die als Leitplanken für den KI-Einsatz dienen. Dazu zählen:

- strategische Zielsetzungen im Hinblick auf KI
- interne Richtlinien und Policies
- Rollen- und Verantwortlichkeitsdefinitionen
- Festlegung akzeptabler Risikoniveaus

Die definierten Maßnahmen müssen schließlich in den operativen Betrieb überführt und in konkrete Handlungsanweisungen und Prozesse übersetzt werden.

Dies erfordert eine enge Abstimmung zwischen strategischer Planung und praktischer Umsetzung.

Trotz der beschriebenen Herausforderungen ist die Auseinandersetzung mit der Operationalisierung von Vertrauenswürdigkeit für Unternehmen unerlässlich. Im folgenden Kapitel wird erläutert, warum sich Organisationen mit dem Thema befassen sollten und welche strategischen Vorteile sich daraus ergeben können.

1.5 Zusammenfassung

Wie in diesem Kapitel aufgezeigt, gibt es starke externe Einflussfaktoren, die Unternehmen dazu motivieren, sicherzustellen, dass die von ihnen entwickelten KI-Systeme

⁵ Vgl. Mittelstadt 2019.

⁶ Vgl. Akbarighatar 2024.

vertrauenswürdig sind. Neben externen Einflussfaktoren können jedoch auch intrinsische Faktoren eine Rolle spielen, um Risiken zu begrenzen und Alleinstellungsmerkmale zu entwickeln. Die konkrete Definition der Vertrauenswürdigkeit eines KI-Systems ergibt sich dabei einerseits aus der Forschung, wie dem KI-Prüfkatalog, regulatorischen Anforderungen, wie

sie im EU AI Act formuliert werden, und dem Werterahmen des jeweiligen Unternehmens. Die Dimensionen, die hierbei betrachtet werden, ähneln sich in der Regel stark, so deckt beispielsweise der Prüfkatalog viele der technischen Themen ab – die im EU AI Act gefordert sind, können jedoch kontextabhängig unterschiedliche Schwerpunkte setzen.



2 Trustworthy AI Operations

In den vorangegangenen Kapiteln wurden die Anforderungen an vertrauenswürdige KI-Systeme analysiert – basierend auf aktuellen Forschungserkenntnissen sowie den regulatorischen Vorgaben, insbesondere im Kontext des EU AI Acts. Dabei wurden zentrale Herausforderungen bei der Operationalisierung identifiziert, etwa die Komplexität des Stakeholdermanagements und die Notwendigkeit einer strukturierten Prozessorganisation zur Etablierung einer effektiven KI-Governance.

Dieses Kapitel widmet sich der Frage, wie sich diese Anforderungen bereits konkret im Entwicklungsprozess von KI-Systemen verankern lassen. Ziel ist es, die Prinzipien vertrauenswürdiger KI nicht nur als externe Anforderungen zu verstehen, sondern als integralen Bestandteil technischer und organisatorischer Entwicklungspraktiken zu etablieren.

Als methodisches Fundament dient das Entwicklungsparadigma der Machine Learning Operations (MLOps), welches die effiziente, skalierbare und nachhaltige Entwicklung von KI-Systemen als Zielsetzung hat.⁷ MLOps verbindet Aspekte der Softwareentwicklung, des Datenmanagements und der Systemintegration und ermöglicht eine kontinuierliche Bereitstellung und Überwachung von KI-Modellen.

Im weiteren Verlauf dieses Kapitels werden zunächst die Grundlagen und Prinzipien von MLOps erläutert. Darauf aufbauend wird dargelegt, wie sich Anforderungen an Vertrauenswürdigkeit systematisch in den MLOps-Prozess integrieren lassen – etwa durch die Erweiterung bestehender Strukturen zur Unterstützung von Risiko- und Qualitätsmanagement. Diese Erweiterung des MLOps-Entwicklungsparadigmas wird als »Trustworthy AI Operations« (TAIOps) bezeichnet. TAIOps verfolgt das Ziel, »Vertrauenswürdigkeit by Design« zu realisieren – also die Berücksichtigung ethischer, sicherheitsrelevanter und regulatorischer Anforderungen bereits in frühen Phasen der KI-Entwicklung.

2.1 Machine Learning Operations (MLOps)

Mit der zunehmenden Professionalisierung der KI-Entwicklung in den letzten Jahren entstand ein dringender Bedarf an strukturierten Entwicklungsparadigmen, die den spezifischen Anforderungen datengetriebener Projekte gerecht werden. Zahlreiche Machine-Learning-Projekte scheiterten am Übergang von der Entwicklung in den produktiven Betrieb und

bleiben im Status experimenteller Prototypen. Eine nachhaltige Produktentwicklung war häufig nicht gegeben.

Die Ursachen hierfür sind vielfältig, lassen sich jedoch in weiten Teilen auf Herausforderungen zurückführen, die auch in der klassischen Softwareentwicklung bekannt sind. Folgerichtig orientierte man sich an den Best Practices aus der Softwareentwicklung und übertrug zentrale Prinzipien und Werkzeuge des dominierenden »Development and Operations«-Paradigmas (DevOps) auf Machine-Learning-Projekte. Diese Übertragung war in vielen Bereichen erfolgreich, erforderte jedoch gezielt Anpassungen, um den besonderen Anforderungen der KI-Entwicklung gerecht zu werden. Insbesondere da KI-Systeme nicht nur aus klassischen Softwarekomponenten wie Schnittstellen, Benutzeroberflächen oder Pipeline-Management-Systemen bestehen, die durch deterministischen Code beschrieben werden können. Sie enthalten zusätzlich eine Modellkomponente, die auf statistischen Zusammenhängen basiert. Die Qualität und das Verhalten dieser Modelle hängen maßgeblich von den Trainings- und Vorhersagedaten ab. Im Gegensatz zu regelbasierten Softwareartefakten modellieren KI-Modelle keine festen Abläufe, sondern lernen Muster und Zusammenhänge aus Daten. Diese Eigenschaft erschwert die Anwendung klassischer Test- und Monitoringverfahren und stellt Entwicklungsteams vor neue Herausforderungen in der Qualitätssicherung und im Betrieb.

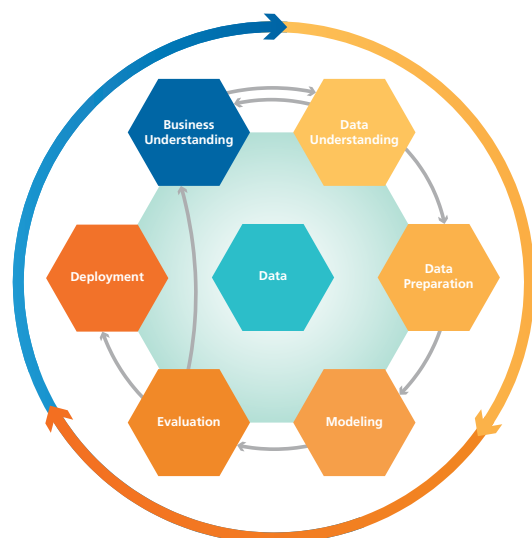


Abbildung 1: CRISP-DM-Vorgehen.⁸

⁷ Vgl. Kreuzberger, Kühl und Hirschl 2023.

⁸ Kenneth Jensen - Own work based on: <ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/18.0/en/ModelerCRISPDm.pdf> (Figure 1), CC BY-SA 3.0, <https://commons.wikimedia.org/w/index.php?curid=24930610>.

Als Antwort auf diese Anforderungen entwickelte sich das Paradigma MLOps. MLOps kombiniert die Prinzipien von DevOps mit spezifischen Verfahren aus dem Bereich des maschinellen Lernens und berücksichtigt dabei den gesamten Lebenszyklus von KI-Modellen. Ein etabliertes Lebensphasenmodell für datengetriebene Projekte bildet seit vielen Jahren der Cross Industry Standard Process for Data Mining (CRISP-DM).

Obwohl CRISP-DM weiterhin eine valide Grundlage bietet, reicht es allein nicht aus, um die Komplexität moderner KI-Systeme vollständig abzubilden. Es fehlen zentrale Elemente wie die Organisation der Zusammenarbeit im Entwicklungsteam und mit Stakeholdern, die Sicherstellung der Systemqualität über alle Komponenten hinweg sowie die Etablierung nachhaltiger Entwicklungsprozesse mit iterativer Anpassungsfähigkeit.

Diese Lücken adressiert das DevOps-Paradigma, das auf konsequente Automatisierung, kontinuierliche Integration und eine enge Verzahnung von Entwicklung und Betrieb setzt. DevOps beschreibt einen zyklischen Prozess, bei dem Erkenntnisse aus dem Betrieb systematisch in die Weiterentwicklung einfließen – ein Prinzip, das sich auch für KI-Projekte als äußerst wertvoll erwiesen hat.

MLOps vereint die Lebensphasenorientierung von CRISP-DM mit den operativen Prinzipien von DevOps und bietet

Entwicklungsteams eine praxisnahe Orientierung für die Entwicklung und den Betrieb von KI-Systemen. Es berücksichtigt die besonderen Anforderungen der KI-Entwicklung – etwa die zentrale Rolle des Datenmanagements, die experimentelle Vorgehensweise bei der Modellerstellung und die interdisziplinäre Zusammensetzung der Teams.

Neben der Definition von Rollen und Prozessen erfolgt eine strukturierte Aufteilung in Entwicklungsphasen. Diese ist insbesondere relevant für die Integration von Vertrauensaspekten: In jeder Phase ergeben sich unterschiedliche Fragestellungen und potenzielle Risiken, die frühzeitig erkannt und adressiert werden müssen. So kann bereits in der Anforderungsanalyse geprüft werden, ob aus dem Projektkontext beispielsweise Datenschutzrisiken entstehen. In der Explorationsphase wiederum lassen sich diese Erkenntnisse nutzen, um beispielsweise gezielt datenschutzkonforme Datenquellen auszuwählen.

Wie eingangs beschrieben, fordert der EU AI Act die Etablierung von Kontrollmechanismen bereits während der Entwicklungsphase. Unternehmen haben ein strategisches Interesse daran, Risiken frühzeitig zu identifizieren und geeignete Maßnahmen zu ergreifen – nicht zuletzt, um Kosten und Zeit zu sparen und die Qualität ihrer KI-Systeme nachhaltig zu sichern.



Abbildung 2: Der MLOps-Zyklus.

2.2 Operationalisierung auf drei Ebenen

Ein MLOps-Prozess, der gezielt darauf ausgerichtet ist, Risiken für die Vertrauenswürdigkeit von KI-Systemen frühzeitig zu identifizieren und zu minimieren, kann einen entscheidenden Beitrag zur Etablierung eines effektiven Risiko- und Qualitätsmanagements leisten. Durch die Integration geeigneter Werkzeuge, Prozesse und Methoden wird es Entwicklungsteams ermöglicht, ihre Arbeit in eine übergeordnete Governance-Struktur einzubetten und vertrauenswürdige KI »by Design« zu entwickeln.

Im Folgenden werden drei zentrale Ebenen beschrieben, die für die Operationalisierung von vertrauenswürdiger KI wesentlich sind: die organisatorische, die prozedurale und die technische Ebene. Jede dieser Ebenen adressiert unterschiedliche Aspekte der Governance und trägt zur systematischen Verankerung von Vertrauenswürdigkeit im Lebenszyklus von KI-Systemen bei.

Organisatorische Ebene

Auf der organisatorischen Ebene werden die übergeordneten Governance-Strukturen definiert, die sicherstellen sollen, dass die eingesetzten Prozesse und Technologien mit den strategischen Zielsetzungen sowie den ethischen Grundsätzen des Unternehmens im Umgang mit KI übereinstimmen.⁹ Dies umfasst die Festlegung von KI-Leitprinzipien, Zielen, Rollen, Verantwortlichkeiten und Rechenschaftsmechanismen, die

sich in den Richtlinien der Organisation widerspiegeln. Solche Governance-Strukturen sind Systeme, »durch die eine Organisation geleitet, beaufsichtigt und für die Erreichung ihres definierten Zwecks verantwortlich gemacht wird.«¹⁰ In diesem Sinne bildet die organisatorische Ebene das Fundament für eine verantwortungsvolle KI-Entwicklung und -Nutzung. Zunächst sollte die Organisation, die KI entwickelt, einsetzt oder nutzt, eine Governance einrichten, um sicherzustellen, dass die technischen und nichttechnischen Anforderungen für den vertrauenswürdigen Einsatz von KI erfüllt sind.

Prozedurale Ebene

Die organisatorischen Vorgaben müssen durch geeignete Prozesse operationalisiert werden, um ihre Ziele im Entwicklungsalltag wirksam umzusetzen. Die ISO 38507 betont, dass sowohl technische als auch nicht-technische Überlegungen bereits zu Beginn eines KI-Projekts berücksichtigt werden sollten – jedoch entwickeln sich diese Anforderungen dynamisch über den gesamten Lebenszyklus hinweg. Daher ist es notwendig, den Entwicklungsprozess durch sogenannte »Quality Gates« zu strukturieren – definierte Punkte im Lebenszyklus, an denen zentrale Fragestellungen wie Risiko- und Qualitätsmanagement systematisch adressiert werden. Diese Gates ermöglichen eine kontinuierliche Reflexion und Anpassung der Maßnahmen und stellen sicher, dass Vertrauenswürdigkeit nicht nur punktuell, sondern phasenübergreifend berücksichtigt wird.

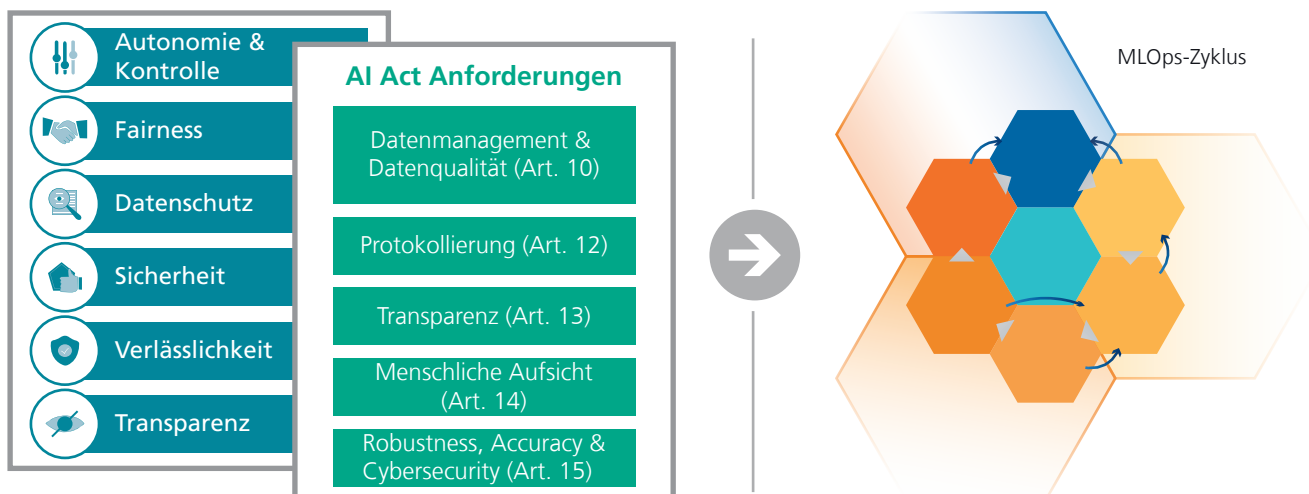


Abbildung 3: Vertrauenswürdige KI »by design«.

⁹ Vgl. Mäntymäki, et al. 2022.

¹⁰ ISO 37000:2021.

Technische Ebene

Auf technischer Ebene erfolgt die konkrete Umsetzung der Prinzipien vertrauenswürdiger KI innerhalb einzelner Systeme. Die in der organisatorischen Ebene definierten Grundsätze werden entlang des KI-Lebenszyklus analysiert und durch die Quality Gates der prozeduralen Ebene strukturiert. Die Bewertung eines KI-Systems erfolgt dabei durch eine risiko-basierte Analyse, die potenzielle Schwachstellen identifiziert und geeignete Gegenmaßnahmen ableitet. Zur Sicherstellung der technischen Qualität kommen Metriken oder Leistungs-indikatoren (KPIs) zum Einsatz – etwa zur Bewertung der Datenqualität, Modellrobustheit oder Fairness. Diese Metriken ermöglichen eine quantifizierbare Bewertung der Wirksamkeit von Maßnahmen und bilden die Grundlage für eine (teil-)auto-matisierte Qualitätssicherung.

2.3 Einbettung von Qualitäts- und Risiko-management in die Entwicklung

TAIOps bezeichnet ein Vorgehensmodell zur Erweiterung von MLOps, das Methoden und Artefakte in die Entwicklung integriert, um die Vertrauenswürdigkeit von KI-Systemen systematisch zu fördern. Während MLOps primär auf Effizienz, Skalierbarkeit und Automatisierung abzielt, erweitert TAIOPs dieses Paradigma um Methodiken zur Förderung der Vertrauenswürdigkeit. Ziel ist es, die Entwicklung und den Betrieb von KI-Systemen so zu gestalten, dass sie nicht nur leistungsfähig, sondern auch verantwortungsvoll und regelkonform sind. Der Ansatz gliedert sich – wie bereits erläutert – in drei miteinander verknüpfte Ebenen: die organisatorische, die

prozedurale und die technische Ebene. Diese Ebenen bilden gemeinsam das Fundament für eine ganzheitliche Operationalisierung vertrauenswürdiger KI.

Auf organisatorischer Ebene werden strategische Zielvorgaben für den vertrauenswürdigen Einsatz von KI definiert. Diese Zielsetzungen besitzen ein hohes Maß an Allgemeingültigkeit und legen fest, wie das Risiko- und Qualitätsmanagement strukturiert sein muss, um für alle Entwicklungsprozesse Gültigkeit zu besitzen. Die Integration in die Unternehmens-Governance erfolgt durch:

- die Definition von KI-Leitprinzipien und ethischen Grundsätzen,
- die Festlegung von Rollen, Verantwortlichkeiten und Rechenschaftsmechanismen,
- die Einführung einer verbindlichen KI-Policy (vgl. Kapitel 3.1).

Diese Governance-Strukturen müssen sicherstellen, dass die eingesetzten Prozesse und Technologien mit den strategischen Zielsetzungen des Unternehmens übereinstimmen. Die konkrete Ausgestaltung ist kontextabhängig und variiert je nach Unternehmensstruktur und Reifegrad. In Kapitel 4 wird anhand einer Fallstudie exemplarisch aufgezeigt, wie eine solche Einbettung gelingen kann. Die prozedurale Ebene dient der Umsetzung der strategischen Zielvorgaben in konkrete Entwicklungsprozesse. Hier werden die MLOps-Prozesse um spezifische Verfahren erweitert, die es ermöglichen, ethische und regulatorische Anforderungen in die tägliche Entwicklungspraxis zu überführen.

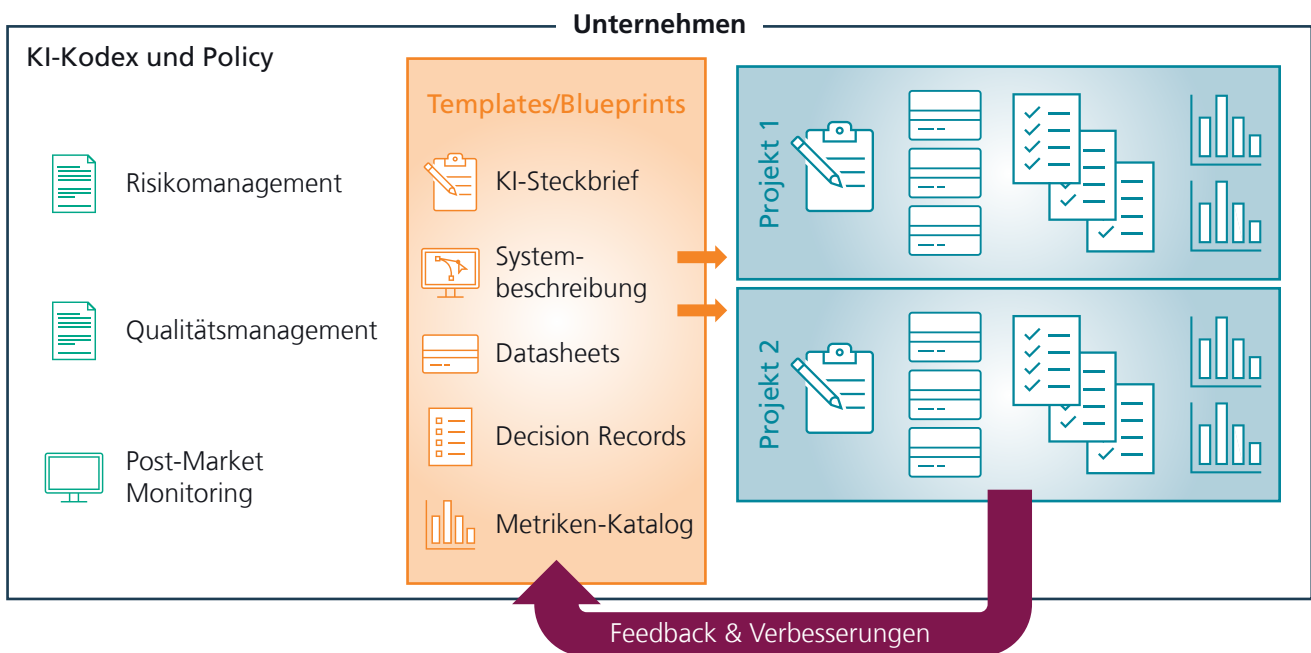


Abbildung 4: Skizze zu unternehmensweiten und projektspezifischen Elementen der TAIOPs-Ebenen.

Die Zielvorgaben der organisatorischen Ebene werden in operative Prozesse übersetzt, indem sie festlegen:

- welche Informationen im KI-Steckbrief (Kapitel 3.2) dokumentiert werden,
- wann Entscheidungen getroffen und in Decision Records protokolliert werden müssen (Kapitel 3.5),
- welche Datenströme relevant sind und durch Datasheets (Kapitel 3.4) beschrieben werden,
- wie die Design- und Architekturdokumentation (Kapitel 3.3) strukturiert wird.

Diese Artefakte sind projektspezifisch und können je nach Anwendungskontext unterschiedliche Schwerpunkte setzen. Die Flexibilität der Umsetzung ist ein zentraler Vorteil des TAI-Ops-Ansatzes: Sie erlaubt es, die konkrete Ausgestaltung an den jeweiligen Anwendungsfall anzupassen, ohne die strategischen Ziele aus dem Blick zu verlieren.

Die prozeduralen Prozesse sind nicht isoliert vom Risiko- und Qualitätsmanagement zu betrachten, sondern dienen als Mittel zur Operationalisierung abstrakter ethischer Zielsetzungen. Sie ermöglichen es, diese Ziele in überprüfbare Maßnahmen zu überführen und tragen somit direkt zur Qualitätssicherung und Risikokontrolle bei.

Auf technischer Ebene erfolgen die konkrete Umsetzung und Überprüfung der Maßnahmen zur Vertrauenswürdigkeit innerhalb einzelner KI-Systeme. Die in der organisatorischen Ebene definierten Grundsätze und die prozeduralen Prozesse bilden die Grundlage für die technische Verifikation.

Hier kommen Methoden und Werkzeuge zum Einsatz, die es ermöglichen

- die Wirksamkeit von Risikokontrollmaßnahmen zu quantifizieren,
- die Einhaltung ethischer Zielsetzungen kontinuierlich zu überwachen,
- die Systemqualität anhand von Metriken und KPIs zu bewerten.

Beispiel hierfür ist der Metrik-Katalog (Kapitel 3.6), aufgegliedert in:

- KI-Testing: Prüfung von Modellen auf Robustheit, Fairness, Sicherheit und weitere Vertrauensdimensionen
- Monitoring: Laufende Überwachung der Systemleistung und Einhaltung definierter Schwellenwerte im Betrieb

Die eingesetzten Metriken und Technologien sind stark kontextabhängig und müssen projektspezifisch gewählt werden. Sie bieten jedoch das höchste Maß an Absicherungsargumentation, da die Zielerreichung mit konkreten Zahlen und Nachweisen belegt werden kann.

Die drei Ebenen von TAI-Ops sind nicht isoliert zu betrachten, sondern bedingen einander. Auf der organisatorischen Ebene werden strategische Zielvorgaben formuliert, die auf der prozeduralen Ebene in operative Prozesse übersetzt und auf der technischen Ebene durch geeignete Metriken überprüft werden.

Ein einseitiger Fokus – etwa auf ethische Zieldefinitionen ohne operative Umsetzung – birgt die Gefahr von »green washing«.¹¹ Umgekehrt führt der ziellose Einsatz technischer Methoden ohne strategische Einbettung zu Frustration und ineffizientem Ressourceneinsatz. Nur durch eine kohärente Umsetzung auf allen drei Ebenen kann Vertrauen in KI-Systeme nachhaltig aufgebaut und regulatorischen Anforderungen entsprochen werden.

2.4 Zusammenfassung

In diesem Kapitel wurde dargestellt, wie sich TAI-Ops als erweitertes Vorgehen im Kontext von MLOps versteht, mit dem Ziel sicherzustellen, dass Methodiken der vertrauenswürdigen KI im Entwicklungsprozess eingebettet werden.

Dazu werden verschiedene Maßnahmen auf organisatorischer, prozeduraler und technischer Ebene umgesetzt. In Bezug auf die Anforderungen des EU AI Acts können so Schritte hin zu einem effektiven Qualitäts- und Risikomanagement unternommen werden.

¹¹ Akbarighatar 2024.

3 TAIOps-Methodik

In den vorangegangenen Kapiteln wurde bereits mehrfach auf Methodiken verwiesen, die zur Operationalisierung von Vertrauenswürdigkeit im KI-Entwicklungsprozess beitragen. Im Folgenden werden diese Instrumente systematisch dargestellt und hinsichtlich ihrer Funktion innerhalb der Governance- und Entwicklungsstruktur erläutert.

3.1 KI-Kodex und -Policy

Zur Schaffung geeigneter Rahmenbedingungen für den verantwortungsvollen Einsatz von KI und zur Etablierung einer transparenten Kommunikation mit internen und externen Stakeholdern greifen viele Unternehmen auf KI-Kodizes und KI-Policies zurück. Im Kern werden in diesen Dokumenten Leitlinien formuliert, die Mitarbeitenden, aber auch Kundinnen und Geschäftspartnern, die Strategie und das Wertesystem des Unternehmens im Hinblick auf die Verwendung von KI vermitteln.

Der KI-Kodex richtet sich primär an externe Stakeholder wie Kunden, Geschäftspartnerinnen oder die Öffentlichkeit. Er formuliert die Werte und Prinzipien, nach denen sich das Unternehmen im Umgang mit KI-Technologien richtet. Kodexe sind in der Regel kurz, abstrakt und öffentlich zugänglich. Sie dienen der Positionierung des Unternehmens und signalisieren dessen ethische Haltung gegenüber KI.

Die KI-Policy hingegen ist ein internes Steuerungsinstrument, das sich an Mitarbeitende und Führungskräfte richtet. Sie definiert die Selbstverpflichtung des Unternehmens im Umgang mit KI und legt fest:

- wie KI-Systeme im Unternehmenskontext eingesetzt werden dürfen,
- welche Schutzbedarfs- und Qualitätskriterien über den gesamten Lebenszyklus hinweg gelten,
- welche Rollen, Verantwortlichkeiten und Prozesse etabliert sind,
- wie die Governance-Strukturen organisiert sind.

KI-Policies sind in der Regel nicht öffentlich, da sie vertrauliche Informationen enthalten und als interne Orientierungshilfe dienen.

Die Vorgaben einer KI-Policy lassen sich durch Handlungsempfehlungen, Guidelines oder Standards weiter spezifizieren. Diese bieten Verantwortlichen eine praxisnahe Orientierung zur Einhaltung der definierten Anforderungen. Beispiele hierfür sind Anleitungen zur Durchführung der Klassifizierung eines KI-Systems nach den Vorgaben des EU AI Acts oder Vorgaben zur Umsetzung und Dokumentation spezifischer Qualitätsanforderungen.

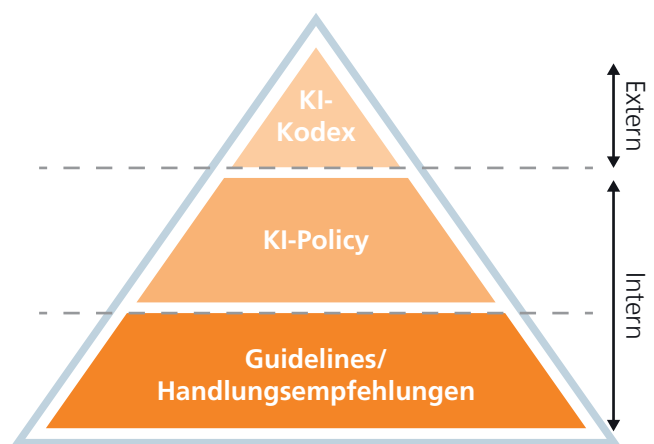


Abbildung 5: Hierarchiepyramide von KI-Kodex, KI-Policy und Handlungsempfehlungen.

3.2 KI-Steckbrief

Ein KI-Steckbrief ist ein zentrales Instrument zur systematischen Dokumentation von KI-Anwendungen. Er wird erstellt, um Interessierten und Betroffenen einen strukturierten Überblick über die Funktionalität, den Einsatzkontext und die zugrundeliegende Technologie eines KI-Systems zu geben.

Im Prüfkatalog des Fraunhofer IAIS wird ein KI-Steckbrief¹² beschrieben und argumentiert, dass dieser die Entwicklung vertrauenswürdiger KI-Lösungen durch eine klare, nachvollziehbare und transparente Dokumentation relevanter Aspekte fördert. Er dient als Grundlage für weiterführende Prüfungen und kann als Kommunikationsmittel zwischen technischen und nicht-technischen Stakeholdern verwendet werden.

¹² Poretschkin, et al. 2021.

Im Folgenden werden zentrale Funktionen eines typischen KI-Steckbriefs dargestellt:

Transparenz und Nachvollziehbarkeit

Der Steckbrief bietet einen kompakten Überblick über die Funktionalität, den Einsatzkontext und die Struktur der KI-Anwendung. Dies schafft Transparenz und ermöglicht es Betroffenen, die grundlegenden Merkmale und Anforderungen der Anwendung zu verstehen.

Frühe Identifikation von Risiken

Durch die strukturierte Beschreibung der grundlegenden Funktionalität und des Einsatzkontextes der KI-Anwendung können potenzielle Risiken und Herausforderungen frühzeitig identifiziert werden. Dies betrifft sowohl regulatorische Anforderungen als auch Aspekte der Wirtschaftlichkeit und des Risikomanagements. Die frühzeitige Erkennung ermöglicht die gezielte Entwicklung geeigneter Maßnahmen.

Förderung einer verantwortungsvollen Nutzung

Die Dokumentation der Schnittstellen zur Umgebung und der menschlichen Interaktion im Betrieb unterstützt eine verantwortungsvolle Nutzung der KI-Anwendung. Ethische Standards und Governance-Vorgaben können somit besser eingehalten und überprüft werden.

Grundlegende Funktionalität und vorgesehener Einsatzkontext (FE)

| Abfragedimension | Beschreibung |
|------------------|---|
| [ST-B-FE-01] | Aufgabe und Funktionalität der KI-Anwendung: Was »macht« sie genau? Welche Eingabedaten und Ausgaben gibt es? |
| [ST-B-FE-02] | Vorgesehener Einsatzkontext: Ist die KI in ein Gesamtsystem eingebettet? Welche Rolle spielen Menschen im Betrieb der KI-Anwendung? |
| [ST-B-FE-03] | Anforderungen an die KI-Anwendung: Regulatorik, Wirtschaftlichkeit, Risikovermeidung. |
| [ST-B-FE-04] | Weitere denkbare Einsatzkontexte und solche, in denen von der Verwendung der KI-Anwendung abgesehen werden sollte. |
| [ST-B-FE-05] | Weitere wichtige Informationen zur Funktionalität oder Betriebsumgebung der KI-Anwendung. |

Struktur der KI-Anwendung (ST)

| Abfragedimension | Beschreibung |
|------------------|--|
| [ST-B-ST-01] | Aufbau der KI-Anwendung: Wichtige Komponenten und deren Funktionalitäten sowie Architektur. |
| [ST-B-ST-02] | Beschreibung der KI-Komponente: Welches ML-Modell wird verwendet? Lernt die KI kontinuierlich? |
| [ST-B-ST-03] | Weitere relevante Punkte zur Struktur der KI-Anwendung. |

Tabelle 1: Übersicht zu den Dimensionen eines KI-Steckbriefes im Sinne des Fraunhofer IAIS Prüfkatalogs.¹³

Dokumentation von Entscheidungen

Der Steckbrief ermöglicht die Nachvollziehbarkeit technischer Entscheidungen, etwa zur Auswahl von Modellen, Architekturen oder Lernverfahren. Dies stärkt die Reproduzierbarkeit und schafft Vertrauen in die Entwicklungsprozesse.

Basis für weiterführende Prüfungen

Der KI-Steckbrief dient als Ausgangspunkt für umfassendere Prüfungen, bei denen detaillierte technische Spezifikationen und Ausführungen verlangt werden.

Der KI-Steckbrief gliedert sich in zwei Hauptdimensionen:

1. Grundlegende Funktionalität und Einsatzkontext (FE)
2. Struktur der KI-Anwendung (ST)

Diese Dimensionen ermöglichen eine ganzheitliche Betrachtung sowohl der operativen als auch der technischen Aspekte der KI-Anwendung.

In Tabelle 1 wird aufgezeigt, welche Dimensionen in einem KI-Steckbrief gemäß den Vorgaben des Prüfkatalogs enthalten sein können.

¹³ Ebd.

Der KI-Steckbrief stellt ein standardisiertes und praxisnahes Werkzeug zur Dokumentation und Bewertung von KI-Anwendungen dar. Er fördert Transparenz, trägt zu einem fundierten Verständnis über die Funktionalität, den Einsatzkontext sowie die Architektur der KI-Anwendung bei und unterstützt die Einhaltung regulatorischer Vorgaben. Als Teil der TAIOPs-Methodik trägt er wesentlich zur Verankerung von Vertrauenswürdigkeit »by Design« bei und schafft eine belastbare Grundlage für Audits, Reviews und interne Qualitätssicherung.

3.3 Systembeschreibung und Architekturdokumentation

Eine umfassende Systembeschreibung und Architekturdokumentation sind zentrale Bestandteile einer erfolgreichen Entwicklung, Implementierung und Wartung von KI-Anwendungen. Sie schafft Transparenz über die Funktionalitäten, Komponenten und eingesetzten Ressourcen eines KI-Systems und bildet die Grundlage für technische Prüfungen, Wartbarkeit und Weiterentwicklung.

Im Unterschied zum KI-Steckbrief ist die Zielsetzung hierbei Prozesse und Datenströme zu visualisieren und den Prozess auf (optimalerweise verschiedenen) Abstraktionsebenen darzustellen. Insbesondere die Darstellung der Datenflüsse durch das System hat im Kontext von KI eine große Bedeutung, aber auch wie das System mit anderen (KI-)Systemen in Beziehung steht.

Zur Darstellung komplexer KI-Systeme hat sich das C4-Modell¹⁴ (Context, Containers, Components, Code) als besonders hilfreich erwiesen. Es bietet eine mehrstufige Herangehensweise zur Visualisierung und Dokumentation technischer Systeme auf vier Ebenen:

1. Context-Diagramm

Zeigt die Einbettung des KI-Systems in seine Umgebung: Welche externen Systeme, Nutzergruppen oder Schnittstellen existieren? Dies ist essenziell für das Verständnis des Einsatzkontextes und der Interaktionen.

2. Container-Diagramm

Stellt die Hauptbestandteile des Systems dar – etwa Microservices, Datenbanken oder Webanwendungen. Es hilft, die Architektur und die eingesetzten Technologien zu überblicken.

3. Component-Diagramm

Geht eine Ebene tiefer und zeigt die internen Komponenten innerhalb der Container. Hier können spezifische Ressourcen

wie Algorithmen, Datenaufbereitungsprozesse oder Evaluationsmethoden dokumentiert werden.

4. Code-Diagramm

Auf der untersten Ebene wird die konkrete Implementierung einzelner Komponenten dargestellt. Dies unterstützt die Nachvollziehbarkeit technischer Entscheidungen und erleichtert die Wartung.

Eine Besonderheit des C4-Modells ist das »Reinzoomen« in immer niedrigere Abstraktionsebenen der KI-Systemarchitektur. Bedeutet, auf der ersten Ebene liegt der Fokus darin zu verstehen, wie das System mit anderen Systemen interagiert und welche Datenströme von außen in das System gelangen, bzw. von diesem erzeugt werden. In den folgenden Diagrammen werden immer detaillierter einzelne Komponenten des Systems dargestellt und für eine immer spezialisiertere Anwendergruppe aufbereitet. So lässt sich das System für Stakeholder mit unterschiedlichsten Hintergründen und Informationsbedarfen auf dem für die Situation angemessenen Detailgrad diskutieren und aufzeigen.

Die Anwendung des C4-Modells ermöglicht die strukturierte, mehrschichtige Dokumentation, die sowohl für technische Teams als auch für Prüfinstanzen verständlich ist, da sie die Zusammenhänge der verschiedenen Systemkomponenten auf verschiedenen »Flughöhen« darstellt. Sie trägt zu mehr Transparenz über die Systemstruktur, Wartbarkeit und Weiterentwicklung von KI-Anwendungen bei und unterstützt somit die Entwicklung vertrauenswürdiger und nachhaltiger KI-Systeme.

3.4 Datasheet

Während der KI-Steckbrief einen strukturierten Überblick über die Funktionalität, den Einsatzkontext und die Struktur einer KI-Anwendung bietet, stellen Datasheets eine vertiefte und systematische Dokumentation der verwendeten Datenressourcen dar. Sie ergänzen den Steckbrief, indem sie spezifische Informationen zu den Daten und den auf diesen basierenden KI-Modellen liefern, dem zentralen Bestandteil jedes KI-Systems.

Datasheets enthalten wesentliche Informationen, die es Stakeholdern ermöglichen, die verwendeten Daten und somit die Stabilität der Modelle zu bewerten. Durch die detaillierte Darstellung von Herkunft, Merkmalen und Anwendungsbereichen tragen Datasheets zur Vertrauensbildung bei und adressieren potenzielle Risiken wie Datenverzerrungen, ethische Bedenken oder Missbrauchspotenziale. Zudem bilden sie den direkten Link zur Data Governance im Unternehmen und stärken die

¹⁴ <https://c4model.com>.

Frage

Für welchen Zweck wurde der Datensatz erzeugt?

Was stellen die Instanzen dar, aus denen sich der Datensatz zusammensetzt (z. B. Dokumente, Fotos, Personen, Länder)?

Gibt es Fehler, Störquellen oder Redundanzen im Datensatz?

Enthält der Datensatz Daten, die als vertraulich angesehen werden könnten (z. B. Daten, die durch das Anwaltsgeheimnis oder die ärztliche Schweigepflicht geschützt sind, Daten, die den Inhalt nicht öffentlicher Kommunikation von Personen enthalten)?

Identifiziert der Datensatz bestimmte Teilpopulationen (z. B. nach Alter, Geschlecht)?

Tabelle 2: Exemplarische Fragen aus »Datasheets for Datasets«¹⁵.

Einbindung der KI-Entwicklung in existierende Governance-Strukturen im Unternehmen.

Datasheets haben in den letzten Jahren große Aufmerksamkeit bekommen und basieren im Wesentlichen auf der Arbeit von Gebru et al.,¹⁵ die, mit dem Ziel, einen Industriestandard zu definieren, über 50 Fragen formuliert haben, die Datensatzersteller bei Veröffentlichung eines neuen Datensatzes beantworten sollen.

Die Einbindung von Datasheets in den MLOps-Prozess erfüllt mehrere zentrale Funktionen. Sie verbessert die Transparenz und Kommunikation, indem sie klar dokumentierte Informationen bereitstellen, die den Austausch zwischen Datenanalysten, Entwicklern und Entscheidungsträgern unterstützen. In stark regulierten Branchen, wie dem Gesundheitswesen oder dem Finanzsektor, unterstützen Datasheets bei der Einhaltung von Datenschutzbestimmungen und ethischen Richtlinien. Eine weitere wesentliche Funktion, die Datasheets im MLOps-Prozess erfüllen, ist Reproduzierbarkeit. Eine präzise Dokumentation der Dateneigenschaften und Modellparameter ermöglicht es, Ergebnisse zu validieren, zu replizieren und weiterzuentwickeln. Dies stellt einen zentralen Aspekt wissenschaftlicher und technischer Integrität dar. Nicht zuletzt stärken Datasheets durch die Offenlegung potenzieller Einschränkungen, Verzerrungen oder

Beschreibung

Gab es eine bestimmte Aufgabe? Gab es eine bestimmte Lücke, die gefüllt werden musste? Bitte beschreiben Sie dies.

Gibt es mehrere Arten von Instanzen (z. B. Filme, Benutzer und Bewertungen; Personen und Interaktionen zwischen ihnen; Knoten und Kanten)? Bitte beschreiben Sie diese.

Wenn ja, geben Sie bitte eine Beschreibung an.

Wenn ja, geben Sie bitte eine Beschreibung an.

Wenn ja, beschreiben Sie bitte, wie diese Teilpopulationen identifiziert werden, und geben Sie eine Beschreibung ihrer jeweiligen Verteilung innerhalb des Datensatzes.

Unsicherheiten das Vertrauen in die verwendeten Daten und die daraus abgeleiteten Modelle.

Zusätzlich zu einer Beschreibung der Daten lässt sich in einem Datasheet darstellen, welche Verarbeitungsschritte in Rahmen des KI-Systems vorgenommen wurden und welche Design Entscheidungen dieses begründen.

3.5 Decision Records

Um Projektteams zu ermöglichen, die Vertrauenswürdigkeit während der Entwicklung zu berücksichtigen, benötigen Entwickler einen strukturierten Ansatz, der sowohl das Bewusstsein für potenzielle Risiken für die Vertrauenswürdigkeit als auch Lösungsstrategien zu deren Minderung umfasst.

Vorab zugewiesene Decision Records stellen eine solche Lösungsstrategie dar, indem sie das Entwicklungsteam dazu motivieren, sich der Risiken für die Vertrauenswürdigkeit bewusst zu werden und diese mit der für die aktuelle Lebenszyklusphase angemessenen Tiefe anzugehen. Zu diesem Zweck werden die relevanten Risikofelder und damit im Zusammenhang stehenden Entscheidungen identifiziert, bspw. durch die Analyse von KI-Prüfkatalogen,¹⁶ und den verschiedenen Lebensphasen des Entwicklungsprozesses zugeordnet. In Abbildung 6 wird hierzu ein Überblick gegeben.

¹⁵ Gebru, et al. 2021.

¹⁶ Vgl. Helmer, et al. 2024.

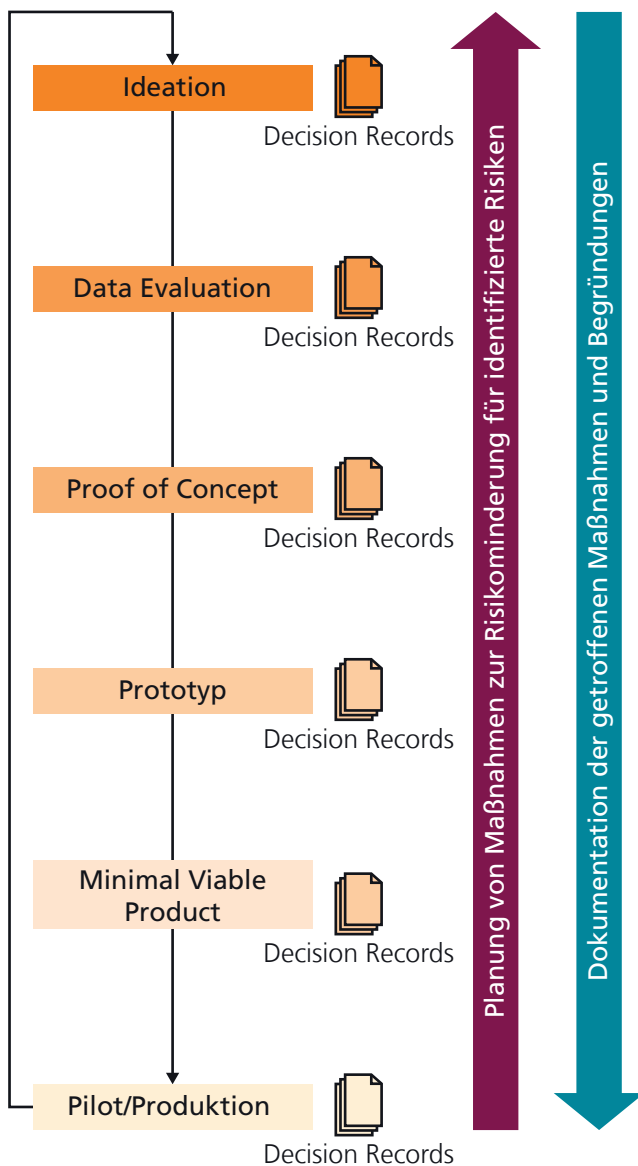


Abbildung 6: Verwendung von Decision Records bei BITMARCK (Kapitel 4).¹⁷

Jeder Decision Record besteht aus zehn Einträgen, wie in Tabelle 3 dargestellt. Die ID dient als eindeutige Kennung für den Decision Record und stellt sicher, dass jeder Eintrag eindeutig referenziert werden kann. Sie ist für die Verfolgung und Organisation mehrerer Datensätze innerhalb einer größeren Datenbank oder eines Dokumentationssystems unerlässlich.

Die Statusbezeichnung gibt den aktuellen Stand der Entscheidung an und kategorisiert sie entweder als angenommen, wie in Tabelle 3 angegeben, offen oder abgelehnt. Diese Klassifizierung ist entscheidend für das Verständnis des Fortschritts des Entscheidungsprozesses und fördert die Transparenz hinsichtlich der Ergebnisse der Beratungen.

Die Identifizierung der Entscheidungsgruppe als die Personen oder Teammitglieder, die am Entscheidungsprozess beteiligt waren, ermöglicht ein Verständnis dafür, wer beteiligt war, und gibt Aufschluss über die Vielfalt der berücksichtigten Perspektiven. Im Beispiel werden John Doe und Jane Smith als Entscheidungsgruppe definiert.

Im Feld »Kontext/Thema« werden Hintergrundinformationen zu der betreffenden Entscheidung bereitgestellt. Dazu gehören kritische Fragen und eine Beschreibung des zu behandelnden Problems. In dem angegebenen Beispiel betont der Text die Bedeutung der Bewertung potenzieller Risiken und Folgen während der Entwurfsphase eines Projekts und geht auf Risiken im Zusammenhang mit personenbezogenen Daten und KI-gestützten Entscheidungen ein. Durch die Bereitstellung des Kontexts hilft diese Kennzeichnung, die Umstände zu verdeutlichen, die die Entscheidung erforderlich gemacht haben, und ermöglicht ein besseres Verständnis ihrer Auswirkungen.

Das Entscheidungslabel hält das endgültige Ergebnis des Beratungsprozesses fest. Es erfasst die konkret getroffene Entscheidung und dient als formelle Dokumentation des von der Entscheidungsgruppe gefassten Beschlusses. Im Beispiel sollte eine gründliche Bias-Analyse der Trainingsdaten und -prozesse sowie eine finanzielle Risikobewertung durchgeführt werden.

Das Feld »Argumente« umreißt die Argumente, die während des Entscheidungsprozesses berücksichtigt wurden. Durch die Dokumentation der Gründe für die Entscheidung bietet es einen ausgewogenen Überblick über die Faktoren, die das Endergebnis beeinflussen, und trägt so zu Transparenz und Rechenschaftspflicht bei. Hier wird ein mittleres Risiko der Diskriminierung von Gruppen aufgrund einer möglichen Verzerrung durch die Verwendung personenbezogener Daten festgestellt, was zu Rechtsstreitigkeiten führen könnte.

Der Abschnitt »Methode« beschreibt die während des Entscheidungsprozesses verwendeten Methoden und Analysewerkzeuge. Er enthält eine Dokumentation der Techniken, die zur Bewertung von Optionen und zur Datenerfassung verwendet wurden, wodurch die Glaubwürdigkeit der Entscheidung erhöht wird. Im Beispiel wurde dieser Punkt kurz durch die Angabe von »Diskussion« als ausgewählte Methode gelöst.

Die Auflistung der relevanten Anforderungen, Standards oder zusätzlichen Entscheidungsaufzeichnungen, die für den Kontext der aktuellen Entscheidung relevant sind, ist wichtig, um sicherzustellen, dass die Entscheidung mit den übergeordneten Organisations- oder Projektzielen übereinstimmt. Im Beispiel werden drei weitere Entscheidungsaufzeichnungen genannt,

¹⁷ Vgl. Helmer, et al., 2024.

| Entscheidungsprotokoll | Beispieltext |
|---|---|
| ID | ID1 |
| Status | Akzeptiert |
| Beteiligte | John Doe, Jane Smith |
| Kontext | <p>Die KI-Lösung sollte keine Personengruppe benachteiligen. Daher ist es unerlässlich, bereits in der Entwurfsphase zu prüfen, ob der Anwendungsfall Risiken birgt, und die Konsequenzen abzuwägen.</p> <p>Leitfragen:</p> <ul style="list-style-type: none"> - Werden personenbezogene Daten verwendet? - Können KI-beeinflusste Entscheidungen zu Benachteiligungen führen? - Was sind die möglichen Folgen fehlgeleiteter Entscheidungen? - Wer ist von solchen Entscheidungen betroffen? |
| Entscheidung | <p>Eine eingehendere Untersuchung der Trainingsdaten, Trainingsprozesse und Modellausgaben auf Verzerrungen ist erforderlich. Zusätzlich muss eine Quantifizierung des finanziellen Risikos durchgeführt werden.</p> |
| Argumente | <p>Das Risiko, bestimmte Gruppen zu benachteiligen, wird als »mittleres Risiko« eingestuft. Die Folgen könnten Klagen mit finanziellen Schäden nach sich ziehen. Der Anwendungsfall erfordert die Nutzung personenbezogener Daten, und es kann nicht ausgeschlossen werden, dass während des Trainings eine Verzerrung erlernt wird.</p> |
| Methodik | Diskussion |
| Verbundene Entscheidungsprotokolle | ID42 (Prüfung der Trainingsdaten), ID24 (Prüfung der Trainingsprozesse), ID21 (Überwachung der Modellausgaben) |
| Verbundene Artefakte | valuation-financial-risk.xls, pitch.ppt |
| Lebensphase | Design |

Tabelle 3: Fiktives Beispiel für einen Decision Record.

nämlich ID42 (Prüfung der Schulungsdaten), ID24 (Prüfung der Schulungsprozesse) und ID21 (Überwachung der Modellausgaben). Wenn diese zu einem späteren Zeitpunkt bearbeitet werden, hängt ihre Bearbeitung ebenfalls von der hier getroffenen Entscheidung ab.

Verwandte Artefakte beziehen sich auf alle ergänzenden Unterlagen oder Artefakte, die während des Entscheidungsprozesses erstellt wurden. Dazu können Berichte, Präsentationen oder Datenanalysen gehören, die weiteren Kontext und Belege für die Entscheidung liefern, wie beispielsweise die im Beispiel aufgeführten Dateien »valuation-fin-risk.xls« und »pitch.ppt«.

Schließlich gibt die Angabe der Lebenszyklusphase den Entwicklungs- oder Fortschrittsstand des Projekts an, auf das sich die Entscheidung bezieht. Diese Klassifizierung hilft, den Kontext der Entscheidung innerhalb der Reifegradstufe des Projekts zu verstehen und liefert Informationen für zukünftige Maßnahmen und Bewertungen.

Da jedes KI-Produkt einen KI-Lebenszyklus durchläuft und die Komplexität eines Projekts während dieses Lebenszyklus zunimmt, müssen verschiedene Arten von Risiken pro Lebenszyklusphase berücksichtigt werden.

Durch die Berücksichtigung und Förderung der Analyse der dringendsten Risiken pro Projektreifegrad ist es möglich, diese ohne unnötige Effizienzverluste anzugehen.

Diese Aufzeichnungen dokumentieren auch Entscheidungen über die Dringlichkeit des Risikos, die zur Minderung ergriffenen Maßnahmen und mögliche Folgeaufgaben.

Darüber hinaus unterstützt dieser Ansatz das Projektteam bei der Kommunikation mit den Stakeholdern, da es über klare Argumente verfügt, welche Risiken auftreten, wie es diese angehen will und welche Ressourcen dafür erforderlich sind.

Nach der Fertigstellung ist es sehr wahrscheinlich, dass die meisten KI-Produkte in Zukunft eine Prüfung bestehen müssen, insbesondere wenn das Produkt aufgrund der EU-Gesetzgebung für den europäischen Markt bestimmt ist. Diese Prüfungen könnten intern von Governance-Abteilungen oder extern von (staatlichen) Behörden durchgeführt werden.

Durch einen strukturierten Prozess, der die Minderung potenzieller Risiken für die Vertrauenswürdigkeit und die Dokumentation relevanter Entscheidungen fördert, sind die Teams und das daraus resultierende Produkt in der optimalen Position, um diese Prüfungen zu bestehen.

3.6 Metrik-Katalog

Das Testen und Überwachen von KI-Systemen während der Entwicklung und im Betrieb ist eine zentrale Tätigkeit zur Sicherstellung der Vertrauenswürdigkeit. Zielsetzung ist es die häufig abstrakten Ziele, wie beispielsweise »Das KI-System ist fair« oder »Das KI-System ist robust«, in quantifizierbare Kennzahlen zu übersetzen, diese zu messen und somit eine robuste Sicherheitsargumentation aufzubauen. Die Herausforderung, mit der sich Entwicklungsteams konfrontiert sehen, liegt in der Identifikation geeigneter Metriken. Hierbei kann ein Katalog geeigneter Metriken, basierend auf den Zielsetzungen, die in der Policy und den Projektzielen formuliert wurden, wichtige Impulse geben und die Vergleichbarkeit projektübergreifend sicherstellen.

Der Katalog sollte dabei direkten Bezug auf gängige Risikoquellen von KI-Systemen nehmen und ggf. auf Design Patterns und Implementierungsanweisungen verweisen, um einen einheitlichen Anwendungskontext und die standardisierte Verwendung sicherzustellen.

Testing

Die Testphase ist ein zentraler Bestandteil im Entwicklungsprozess von KI-Systemen. Sie stellt sicher, dass die KI-Modelle nicht nur korrekt funktionieren, sondern auch robust und zuverlässig sind. Wesentliche Aspekte des KI-Testings betreffen die Auswahl der verfügbaren Testmethoden, die Bedeutung

von geeigneten Testdaten und die Herausforderungen, die bei der Validierung von KI-Systemen auftreten können.

Testmethoden

Beim Testen von KI-Systemen kommen verschiedene Methoden zum Einsatz, um die Leistungsfähigkeit, Genauigkeit und Performanz der Modelle zu überprüfen. Zu den gängigen Testmethoden gehören:

- **Unit-Tests**

Diese Tests überprüfen einzelne Komponenten oder Module des KI-Systems, um sicherzustellen, dass sie wie erwartet funktionieren.

- **Integrationstests**

Diese Tests stellen sicher, dass die verschiedenen Komponenten des KI-Systems nahtlos zusammenarbeiten.

- **Systemtests**

Diese Tests bewerten das gesamte KI-System in einer realistischen Umgebung, um sicherzustellen, dass es die gewünschten Ergebnisse liefert.

- **Benchmarks**

Insbesondere generative Modelle werden durch Benchmark-Datensätze getestet. In diesen (mitunter auch öffentlich verfügbaren) Datensätzen werden beispielsweise Aufgaben mit Bezug zu Geschlechtern gestellt und geprüft, ob das Modell diese angemessen und ohne Bias löst.

Testdaten

Die Qualität der Testdaten ist entscheidend für die Aussagekraft der Testergebnisse. Die Testdaten sollten repräsentativ für die realen Daten sein, mit denen das KI-System im Betrieb konfrontiert wird. Darüber hinaus sollten die Testdaten vielfältig und umfassend sein, um sicherzustellen, dass das KI-System in der Lage ist, eine breite Palette von Szenarien zu bewältigen. Auch das ergänzende Hinzufügen von Daten, die zunächst »out-of-distribution« sind, ist im Hinblick auf die Robustheit des Systems durchaus sinnvoll. Die Dokumentation der Auswahl und Zusammensetzung der Testdaten im Datasheet (Kapitel 3.4) stellt sicher, dass Tests robust sind und wiederholt werden können.

Herausforderungen beim KI-Testing

Das Testen von KI-Systemen bringt spezifische Herausforderungen mit sich. Eine der größten Herausforderungen besteht darin sicherzustellen, dass KI-Modelle nicht nur auf Trainingsdaten performant sind, sondern auch auf neuen, unbekannteren Daten präzise Vorhersagen treffen. Probleme wie »Training-Serving Skews«, »Overfitting« oder des Generalisierungsfehlers werden zwar in der Ausbildung von Datenwissenschaftlerinnen und -analysten ausführlich diskutiert, stellen aber nach wie vor Hürden dar, die es in konkreten Projektvorhaben zu überwinden gilt. Darüber hinaus muss sichergestellt werden,

dass KI-Modelle fair und unvoreingenommen sind, um Diskriminierung zu vermeiden.

Testing als Enablement für Monitoring

Sobald man Metriken hat, anhand derer man KI-Systeme auf Vertrauenswürdigkeit testen kann, ist auch die Anschlussverwendung dieser Metriken als Teil des Monitorings grundsätzlich möglich. Interessant sind hierfür vorwiegend Metriken, die zwar inhaltliche Aufschlüsse auf die Ergebnisse des Modells geben sollen, aber dennoch auch routinemäßig im Betrieb erhoben werden können. Zu diesem Zweck sind groß angelegte Benchmarks in den seltensten Fällen heranzuziehen. Weitere Aspekte des Monitorings werden im folgenden Abschnitt diskutiert.

Monitoring

Das Monitoring von KI-Systemen ist ein kontinuierlicher Prozess, der sicherstellt, dass KI-Modelle auch nach ihrer Bereitstellung zuverlässig und effizient funktionieren. Wichtige Aspekte des Monitorings von KI-Systemen beinhalten die Überwachung der Modellleistung, die Identifizierung von Anomalien, aber auch das (manuelle oder automatische) Auslösen regelmäßiger Aktualisierungen. Im Falle manueller Aktualisierungen spielt das »Alerting« – also das begründete Auslösen eines Alarms oder Benachrichtigung – eine zentrale Rolle.

Überwachung der Modellleistung

Die Leistung von KI-Modellen kann sich im Laufe der Zeit ändern, insbesondere wenn sich die zugrunde liegenden Daten oder die Betriebsumgebung ändern. Daher ist es wichtig, die Modellleistung kontinuierlich auf einem aktualisierten Datenbestand zu überwachen, um sicherzustellen, dass die Modelle weiterhin genaue und zuverlässige Vorhersagen treffen. Zu den gängigen Methoden zur Überwachung der Modellleistung gehören:

- **Leistungsmetriken**

Regelmäßige Auswertung von Metriken wie Präzision, Recall und F1-Score, um die Modellleistung zu bewerten.

- **Drift-Erkennung**

Identifikation von Daten- und Konzeptdrift, also Veränderungen in den Datenverteilung oder im Zusammenhang zwischen Eingabe und Ausgabe. Hierzu sind *statistische Tests* verwendbar – auch im Fall, dass keine »gelabelten« Daten im Betrieb erfasst werden können. Etablierte statistische Tests, die hierzu geeignet sind und schnell implementiert werden können, sind der Kolmogorov-Smirnov-Test für numerische Features und der Chi-square-Test für kategorische Features.

Anomalie-Erkennung

Anomalien in den Daten oder im Verhalten des KI-Systems können auf Fehlfunktionen, Sicherheitsrisiken oder ethische

Probleme hinweisen. Durch die Implementierung von Anomalie-Erkennungsverfahren können ungewöhnliche Muster oder Abweichungen frühzeitig erkannt und entsprechende Maßnahmen ergriffen werden. Dies kann dazu beitragen, die Zuverlässigkeit und Sicherheit des KI-Systems zu gewährleisten. Die gängigsten Verfahren zur Anomalie-Detektion sind auch heute noch statistische Tests (ähnlich wie bei der Drift-Erkennung), aber auch eine Reihe von (klassischen) Verfahren des maschinellen Lernens wie Decision Trees, k-Nearest Neighbors oder andere Clustering-Verfahren können Aufschluss über Anomalien geben.

Bedeutung regelmäßiger Aktualisierungen

KI-Modelle müssen regelmäßig aktualisiert werden, um ihre Leistung und Relevanz zu erhalten. Dies kann durch erneutes Training auf einem aktualisierten Datenbestand (»Retraining«) geschehen, durch Anpassung der Modellparameter oder in seltenen Fällen durch Implementierung neuer Algorithmen bzw. Modell-Architekturen erfolgen. Regelmäßige Aktualisierungen stellen sicher, dass die Modelle auf dem neuesten Stand bleiben und weiterhin den Anforderungen der Benutzer entsprechen. Um festzustellen, wann eine Aktualisierung erforderlich ist, müssen Kenngrößen und Metriken erfasst werden, für die das Über- oder Unterschreiten eines Schwellwerts einen Alarm (»Alert«), einen Retraining-Prozess oder die Verarbeitung bzw. Anbindung weiterer Daten zur Verbesserung des Gesamtsystems auslöst.

3.7 Zusammenfassung

In Kapitel 3 wurden verschiedene Methodiken beschrieben, die in den KI-Entwicklungsprozess integriert werden können, um sicherzustellen, dass die notwendigen Prozesse und Dokumentationen aufgebaut werden.

Auf organisatorischer Ebene wird in einem KI-Kodex das Wertegerüst definiert, anhand dessen im Unternehmen KI entwickelt und betrieben wird, und in einer KI-Policy die Verantwortlichkeiten, Rollen und Prozesse beschrieben, nach denen sich Mitarbeitende richten können.

KI-Steckbrief, Datasheet, standardisierte Systembeschreibungen und Entscheidungsprotokolle können eingesetzt werden um Dokumentation, Transparenz und Nachvollziehbarkeit auch nach Projektende sicherzustellen und Mitarbeitende und Entwickler dabei unterstützen, Risiken zu identifizieren, zu dokumentieren und angemessene Maßnahmen zu ergreifen.

Schlussendlich wird die Bedeutung von KI-spezifischem Testen und Monitoring aufgezeigt, um auch im Betrieb die Vertrauenswürdigkeit sicherstellen zu können.

4 BITMARCK – KI-Governance für eine vertrauenswürdige KI

von Sermad Abbas, Data Scientist, BITMARCK

In diesem Kapitel wird anhand eines praxisnahen Use Cases dargelegt, wie ein Unternehmen bei der Einführung und Weiterentwicklung von TAIOps in Kooperation mit dem Fraunhofer IAIS vorgegangen ist. BITMARCK beschäftigt sich bereits seit einigen Jahren mit der KI-Entwicklung und damit, wie sichergestellt werden kann, dass die entwickelten Systeme vertrauenswürdig und gesetzeskonform sind. Damit ist das Unternehmen ein praxisnahes Beispiel, wie KI-Governance-Prinzipien erfolgreich in die Unternehmensstrategie und Entwicklungsprozesse eingebunden werden können und bietet wertvolle Einblicke in die Erfahrungen der Entwickler und Prozessverantwortlichen.

Als führender Digitalisierungspartner der gesetzlichen Krankenversicherung treibt BITMARCK die digitale Transformation in der Branche mit innovativen Produkten und Services voran. Grundlage hierfür ist der GKV-Softwarestandard BITMARCK_21cJng, der bei den angeschlossenen Krankenkassen im Einsatz ist. Kunden der Unternehmensgruppe sind die Betriebs- und Innungskrankenkassen sowie die DAK-Gesundheit und weitere Ersatzkassen – über 30.000 Mitarbeitende und rund 35 Millionen Versicherte in der GKV profitieren von den IT-Dienstleistungen von BITMARCK, mehr als 80 Prozent der deutschen gesetzlichen Krankenkassen sind Kunden der Unternehmensgruppe. Mit mehr als 1.900 Mitarbeitenden erzielt BITMARCK einen Jahresumsatz von mehr als einer halben Milliarde Euro.

Die Arbeit von BITMARCK bewegt sich in einem durch den Gesetzgeber stark regulierten Umfeld. Die im Auftrag der gesetzlichen Krankenkassen verarbeiteten Sozialdaten sind als besonders schützenswert eingestuft und dürfen nur für die im Sozialgesetzbuch genannten Zwecke verwendet werden. Dies bedingt, dass die Einhaltung der gesetzlichen Vorgaben stets sichergestellt sein muss und die Verarbeitung der Daten hohen Sicherheitsansprüchen genügt.

Durch die Entwicklung des GKV-Softwarestandards BITMARCK_21cJng ist bei BITMARCK eine große Expertise im Bereich der traditionellen Softwareentwicklung nach agilen Prinzipien vorhanden. Aufgrund der steigenden Anzahl an KI-Projekten mussten die bestehenden Entwicklungsstandards erweitert werden, um den in Kapitel 3.1 genannten Anforderungen gerecht zu werden.

KI-Projekte bei BITMARCK bewegen sich von prädiktiven Modellen bis hin zu generativen KI-Modellen zur

Prozessautomatisierung und -unterstützung, mit dem Ziel, Mitarbeitende in den Krankenkassen zielgerichtet zu entlasten und die Versorgungsqualität der Versicherten zu erhöhen.

Um den Erwartungen der Krankenkassen und den Versicherten an einen vertrauenswürdigen Umgang mit den sensiblen Sozialdaten gerecht zu werden, wurde unter Maßgabe einer schnellen, sicheren und standardisierten KI-Entwicklung beschlossen, KI-Systeme von Beginn an nach den vorgestellten Methodiken zu entwickeln. In den folgenden Kapiteln wird beschrieben, wie BITMARCK die TAIOps-Methodik praxisnah ausgestaltet hat.

4.1 Das Lebensphasenmodell

Der erste Schritt zur Standardisierung der KI-Entwicklung war die Einführung des MLOps-Paradigmas für die strukturierte Bearbeitung von KI-Projekten. Basierend auf dem MLOps-Zyklus wurde der Rahmen für verschiedene Aktivitäten zur Formalisierung des KI-Entwicklungsprozesses gebildet. Das Ziel war eine qualitativ hochwertige und effiziente Entwicklung von KI-Systemen unter Einhaltung regulatorischer Richtlinien. Zu diesem Zweck wurde aus dem MLOps-Zyklus ein Lebensphasenmodell abgeleitet, das alle organisatorischen, prozeduralen und technischen Aspekte in der Entwicklung berücksichtigt. Außerdem wurde eine Entwicklungsplattform für die sichere, skalierbare Entwicklung und Bereitstellung von KI-Systemen aufgebaut, mit der die MLOps-Prinzipien durch die Verwendung entsprechender Tools im Entwicklungsalltag integriert werden können.

Das Lebensphasenmodell wurde ausgehend von Erfahrungen aus vorherigen KI-Projekten erarbeitet und dient sowohl als Unterstützung für die KI-Entwicklung als auch als Kommunikationsmedium gegenüber den Stakeholdern eines Projekts. Es gibt den Entwicklungsteams Orientierung, indem es Aspekte betont, die zu einem bestimmten Zeitpunkt in der Entwicklung im Fokus stehen und definiert Qualitätskriterien, die für den Übergang von einer Lebensphase in die nächste erfüllt sein müssen. Somit kennen Entwicklungsteams die formalen Anforderungen und haben Klarheit darüber, worauf sie sich in einer Projektphase konzentrieren müssen.

Die Qualitätskriterien sind auch Entscheidungsgrundlage für die Stakeholder, denen damit der Zustand eines Projekts

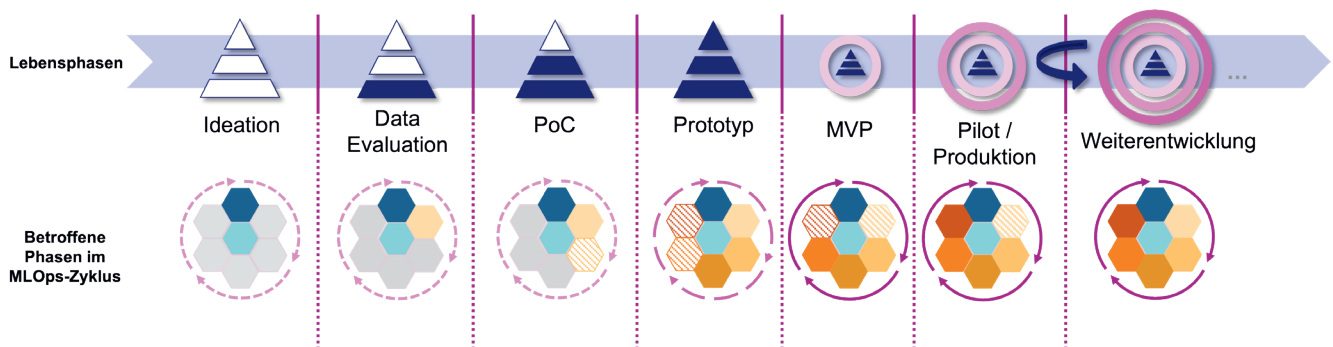


Abbildung 7: Zuordnung der MLOps-Phasen zu den Lebensphasen im BITMARCK KI-Lebensphasenmodell.

transparent dargestellt wird. Im Sinne eines Fail-Fast-Ansatzes können Projekte frühzeitig beendet werden, wenn sich abzeichnet, dass die beabsichtigte Produktionsreife oder Projektziele nicht erreicht werden können.

Abbildung 7 zeigt, wie die MLOps-Prozessschritte aus Abbildung 2 dem BITMARCK-Lebensphasenmodell für KI-Projekte zugeordnet werden.

In jeder Lebensphase wird ein Teil des MLOps-Prozesses durchlaufen. Während das Lebensphasenmodell geradlinig verläuft und keine Rückkehr in eine vorherige Lebensphase vorsieht, kann innerhalb einer Lebensphase durchaus in eine vorangehende MLOps-Phase zurückgekehrt werden, sollte dies nötig sein.

4.2 Umsetzung von TAIOps bei BITMARCK

Als Vorbereitung auf die Einführung des EU AI Acts wurde bei BITMARCK frühzeitig damit begonnen, die Dimensionen der Vertrauenswürdigkeit im Lebensphasenmodell zu berücksichtigen. Daraus entstand mit dem TAIOps-Leitfaden eine Anleitung für die Entwicklungsteams. Der Leitfaden steht im Mittelpunkt des unternehmensweiten KI-Qualitätsmanagementsystems und beinhaltet Vorlagen für den KI-Steckbrief, das C4-Modell, das Datasheet und die Decision Records. Ergänzt wird der TAIOps-Leitfaden durch eine Checkliste zur Prüfung der Einhaltung von Qualitätskriterien, die beim Phasenübergang erfüllt sein müssen.

Die Zusammenstellung des TAIOps-Leitfadens und die Anpassung der einzelnen Dokumente an die BITMARCK-Landschaft erfolgte interdisziplinär, um die regulatorische Perspektive und die Entwicklungssicht abzubilden. Damit wurde sichergestellt, dass die gewonnenen Erkenntnisse den regulatorischen Anforderungen genügen und sich stimmig in das Lebensphasenmodell einfügen, um den Zusatzaufwand für die Entwicklungsteams möglichst gering zu halten.

Sigrun Schnurbusch-Grund, Expertin für Data und AI Governance, beschreibt den TAIOps-Leitfaden als Grundlage für die Etablierung einer unternehmensweiten KI-Governance.



Gerade im Gesundheitswesen ist es wichtig, dass alle KI-Systeme sicher, transparent und gesetzeskonform entwickelt und betrieben werden. Wenn wir ein Qualitätsmanagementsystem einführen, dann nicht nur, weil es gesetzlich gefordert ist, sondern weil wir selbst einen hohen Anspruch an vertrauenswürdige KI haben. Dabei ist essenziell, den konkreten Nutzen für jeden Einzelnen im Arbeitsalltag sichtbar zu machen. Nur so fällt es leichter, neue Arbeitsweisen zu übernehmen und alte Gewohnheiten loszulassen. Die Herausforderung liegt darin, Standards, Strukturen und Prozesse in einer Art und Weise zu gestalten, dass sie nicht nur zusätzliche Dokumentation bedeuten, sondern vor allem Klarheit und Vereinfachung im Rahmen der KI-Entwicklung bringen. Der TAIOps-Leitfaden bildet dabei den Kern unseres Qualitätsmanagementsystems und soll genau das sicherstellen.«

Sigrun Schnurbusch-Grund
Expertin für Data und AI Governance

Der KI-Steckbrief

Der KI-Steckbrief ist das zentrale Kommunikationsmedium für eine transparente und übersichtliche Darstellung des KI-Systems. Sowohl Stakeholder außerhalb des Teams als auch neue Teammitglieder bekommen einen schnellen Überblick über die Inhalte und die wesentlichen Entscheidungen, die im Projektverlauf getroffen wurden.

Der Steckbrief wird über den Lebenszyklus eines KI-Systems gepflegt, damit er stets eine zuverlässige Informationsquelle darstellt. Zu Beginn enthält er rudimentäre Informationen zum Projekt, die im Verlauf immer weiter ausdetailliert

werden. Das Entwicklungsteam legt darin den Aufgabenbereich, die Methodik und den Betrieb des KI-Systems dar. Um zu vermeiden, dass Teams gleiche Dokumentation an verschiedenen Stellen nachhalten müssen, ist ein Verweis auf bestehende Dokumente möglich, ohne den Inhalt explizit im KI-Steckbrief zu wiederholen.

Teile des Steckbriefs werden für die unternehmensinterne Inventarisierung aller entwickelten oder sich in Entwicklung befindlichen KI-Systeme mitsamt Risikoklassifikation gemäß EU AI Act verwendet. Er bildet damit auch eine Brücke zur unternehmensweiten KI-Governance.

Das C4-Modell

Die Nutzung des C4-Modells bei BITMARCK folgt der Beschreibung in Kapitel 3.3.

In den frühen Lebensphasen liegt der Fokus auf dem Context-Diagramm. Dieser geringe Detailgrad gibt den Entwicklungsteams die Möglichkeit, bereits früh mit den Stakeholdern verschiedene Wege für die Integration des KI-Systems in die Gesamtarchitektur zu diskutieren, ohne sich in Feinheiten zu verlieren.

Die zusätzlichen Informationen, die das Projektteam während der Entwicklung erhält, werden genutzt, um die Systemarchitektur in den folgenden Ebenen weiter zu verfeinern.

Das C4-Modell dient den Teams somit nicht nur als reines Dokumentationswerkzeug, sondern unterstützt auch bei der Erarbeitung von Lösungen, indem verschiedene Optionen sowohl im Team als auch mit den Stakeholdern diskutiert werden. Die visuelle Darstellung erleichtert dies.

Vorgeschrieben ist bei BITMARCK lediglich eine Erstellung von Diagrammen der ersten drei Ebenen, also des Context-, des Container- und des Component-Diagramms. Teams können nach eigenem Ermessen entscheiden, ob sie auch ein Code-Diagramm erstellen.

Das Datasheet

Der Umgang mit sensiblen Daten und Nachweispflichten, wie Daten verwendet werden, machte es für BITMARCK notwendig, abweichend von der in Kapitel 3.4 vorgestellten Struktur, ein eigenes Datasheet zu entwickeln, das sich in die unternehmensinterne Data Governance einfügt.

Statt der Verwendung standardisierter Fragenkataloge, werden die Entwicklungsteams durch verschiedene Bereiche mit fest vorgegebener Antwortstruktur, zum Beispiel durch auszuwählende Checkboxen oder auszufüllende Tabellen, und nur geringem Freitextanteil geführt. In Abstimmung mit den Abteilungen für Datenschutz und Informationssicherheit wurde somit ein Dokument geschaffen, das sowohl regulatorischen

Anforderungen genügt als auch leicht in den Arbeitsalltag der Entwicklungsteams zu integrieren ist.

Die Datenschutz- und Informationssicherheitsbeauftragten können somit zielgerichtet die benötigten Informationen einsehen. Gleichzeitig wird durch die feste Struktur eine standardisierte Darstellung geschaffen, die den Entwicklungsteams das Nachhalten der Datenverarbeitungsaktivitäten erleichtert, da der Aufwand hierfür deutlich reduziert wird.

Wie auch KI-Steckbrief und C4-Modell wird das Datasheet laufend gepflegt, um einen aktuellen Stand zu gewährleisten. Die vorgegebene Antwortstruktur vereinfacht nachträgliche Änderungen.

Das bei BITMARCK verwendete Datasheet zeichnet sich vor allem dadurch aus, dass es eine Verbindung zwischen den Daten und einer spezifischen Verarbeitungsaktivität gibt. Für die Nachvollziehbarkeit einzelner Verarbeitungsaktivitäten wird festgehalten, von welchem System auf welche Daten zugegriffen wird. Zudem wird notiert, ob diese Aktivität relevant für eine Prüfung durch den Datenschutz oder die Informationssicherheit ist.

Decision Records

Für jede Lebensphase des KI-Entwicklungszyklus wurden Decision Records (siehe Kapitel 3.5) definiert. Diese Zuordnung ermöglicht den Entwicklungsteams eine gezielte Auseinandersetzung mit den zum Entwicklungszeitpunkt relevanten Fragestellungen und verhindert gleichzeitig eine Überforderung durch vorzeitige Komplexität.

Die Bearbeitung der Decision Records erfolgt zu Beginn einer Lebensphase. Wenn Maßnahmen zu ergreifen sind, werden Einträge im Product Backlog erstellt. Die Umsetzungsschritte werden anschließend vom Team in die Sprints integriert.

Product Owner Daniel Cebe bewertet die Decision Records als *»besonders hilfreich im Rahmen einer vertrauenswürdigen KI«*. Sie unterstützen die Teams dabei, sich *»mit allen Facetten der bekannten sechs Dimensionen konkret [zu] beschäftigen und diese entsprechend berücksichtigen [zu] können«*.

Die Decision Records bauen zum Teil aufeinander auf. Die Teams sind daher angehalten, sich mit bestimmten Themen in verschiedenen Lebensphasen auseinanderzusetzen. Wurde beispielsweise in einer vorherigen Lebensphase bereits festgestellt, dass eine Thematik für das Produkt nicht relevant ist, so kann auf diese Erkenntnisse verwiesen werden. Genauso führt dies aber auch dazu, dass getroffene Entscheidungen mit den weiteren Informationen, die sich im fortschreitenden Projektverlauf ergeben haben, hinterfragt, angepasst und angereichert werden können.

Die Decision Records sind somit nicht nur ein Werkzeug, um in einem Audit nachweisen zu können, dass relevante regulatorische Aspekte berücksichtigt wurden, sondern dienen auch den Entwicklungsteams selbst als Hilfsmittel, um vergangene Entscheidungen nachzuvollziehen und daraus zukünftige Maßnahmen ableiten zu können.

Data Scientistin Lisa Prepens beschreibt die ersten Erfahrungen mit den Decision Records als anfangs »*enorm viel Arbeit, weil viele Fragestellungen zum ersten Mal auftauchen*« und nicht immer ein direkter Projektbezug erkennbar gewesen sei. Trotzdem wird der Nutzen positiv betrachtet. »*Nach der schwerfälligen Erfahrung in der ersten Projektphase waren die Decision Records in der nächsten Lebensphase für uns einfacher händelbar. Zum einen, weil wir mit der Vorgehensweise und den grundlegenden Inhalten schon vertraut waren. Zum anderen, weil das meiste tatsächlich auf den vorhergehenden Überlegungen aufbaut.*«

Qualitätskriterien für den Phasenübergang

Der Abschluss einer Projektphase wird anhand verschiedener Qualitätskriterien definiert. Hierzu wurde eine Checkliste mit Mindestanforderungen erstellt, die für den Übergang in die nächste Projektphase erfüllt sein müssen. Sie beziehen sich sowohl auf den Ausfüllgrad der vorgenannten Dokumente, beinhalten aber auch Aspekte zur technischen Umsetzung, wie zum Beispiel Voraussetzungen an den Implementierungsfortschritt und die durch Kennzahlen quantifizierte Qualität des KI-Systems.

Die Anforderungen sind so gestaltet, dass für die Teams übersichtlich dargestellt ist, bis zu welchem Detailgrad welche Arbeiten am System zu einem bestimmten Zeitpunkt erfüllt sein müssen. Sie halten die Entwicklungsteams aber nicht davon ab, mehr umzusetzen als gefordert ist.

4.3 Einführung des TAIOps-Leitfadens in den Projekten

Die Einführung des TAIOps-Leitfadens erfolgte iterativ. Begonnen wurde mit einer Pilotierung in einem Team, das sich in der Ideation-Phase, der ersten Lebensphase eines Projekts, befand.

Das Team wurde zu Beginn mit den Anforderungen des EU AI Acts vertraut gemacht und damit, wie der TAIOps-Leitfaden dabei unterstützen kann, diesen Anforderungen gerecht zu werden. In der Anfangsphase erfolgte eine enge Begleitung durch Kolleginnen und Kollegen der Arbeitsgruppe, die den Leitfaden entwickelt hat. Regelmäßige Austauschtermine gaben die Möglichkeit, Rückmeldungen zum Leitfaden zu geben. Dies ermöglichte eine gezielte Unterstützung des

Teams sowie eine Anforderungsaufnahme für Verbesserungen in zukünftigen Iterationen des Leitfadens.

Mit Eintritt in die nächste Lebensphase erfolgte die Arbeit innerhalb des Teams selbstständig, wobei stets eine Rückkopplung zur Arbeitsgruppe bestand.

Nach ähnlichem Prinzip wurde der TAIOps-Leitfaden in weiteren Teams eingeführt, um die Arbeitsweise mit ihm zu verstetigen und ihn zu einer Selbstverständlichkeit während der KI-Produktentwicklung zu machen.

Parallel zur Nutzung in den Projekten wird der Leitfaden anhand der Rückmeldungen aus den Teams konsequent weiterentwickelt, um sowohl die Arbeit damit als auch die Zugänglichkeit zu vereinfachen.

4.4 Erkenntnisse und Empfehlungen

Basierend auf den Erfahrungen der Nutzenden lassen sich folgende Empfehlungen für die Einführung und die Arbeit mit dem TAIOps-Leitfaden hervorheben:

- Die Begleitung in der Anfangsphase hat sich als sinnvoll herausgestellt, um den Teams die Arbeit mit dem kompletten Leitfaden, der auf den ersten Blick nach hohem zusätzlichem Aufwand aussieht, zu erleichtern. Insbesondere konnte so noch einmal Klarheit über die Bedeutung der einzelnen Komponenten des Leitfadens für die Entwicklung eines KI-Systems geschaffen werden.
- Regelmäßige Austauschtermine zur Besprechung von Unklarheiten und Verbesserungsvorschlägen helfen, den Leitfaden mehr an die konkreten Bedürfnisse der Teams anzupassen, sodass er sich gut in den Entwicklungsprozess einfügt und nicht als zusätzliche Belastung wahrgenommen wird. Die Akzeptanz in den Teams wird dadurch erhöht. In diesem Zuge ist auch die Einführung eines strukturierter Rückmeldewegs hilfreich, um Anregungen der Teams zu erfassen, die in zukünftigen Versionen des Leitfadens berücksichtigt werden können.
- Eine gute Kommunikation stellt sicher, dass der Leitfaden von allen Entwicklungsteams wahrgenommen wird. Hierzu gehört nicht nur die Bekanntmachung, dass er als Dokumentationswerkzeug existiert, sondern eine ausführliche Einführung am Anfang eines jeden Projekts.
- Generell sollte darauf geachtet werden, dass die Bearbeitung der Komponenten KI-Steckbrief, Systemarchitektur, Datasheet und Decision Records zeitnah erfolgt und nicht erst kurz vor Abschluss einer Lebensphase. Abhängig vom jeweiligen Projektstand kann eine solche rückwirkende

Arbeit mit einem erheblichen Mehraufwand für das Entwicklungsteam verbunden sein. Insbesondere im Fall der Decision Records besteht die Gefahr, dass sie nur genutzt werden, um getroffene Entscheidungen nachträglich zu rechtfertigen, weil die intensive Auseinandersetzung mit den Aspekten eines vertrauenswürdigen KI-Systems und die Ableitung entsprechender Maßnahmen einen hohen Modifikationsaufwand bedeuten kann, der zu einer Verzögerung der Entwicklung führt. Genauso bietet auch die frühzeitige Erstellung eines Architekturbilds die Gelegenheit, verschiedene Lösungswege zu durchdenken und zu einer gut begründeten Lösung zu kommen, welche dann strukturiert umgesetzt werden kann.

- Es sollte sichergestellt sein, dass alle relevanten Stakeholder und die Entwicklungsteams ein einheitliches Verständnis der Begrifflichkeiten haben und eine konsistente Kommunikation in Richtung der Kunden, des Managements und des Entwicklungsteams erfolgt. Nur so kann garantiert werden,

dass gleiche Erwartungen zur Entwicklung auf allen Ebenen vorliegen und es keine unrealistischen Zeitpläne gibt, die die Teams dazu drängen Aufgaben weniger intensiv zu bearbeiten. Spätestens in einem Audit könnte dies nämlich zu aufwändigen Nacharbeiten oder auch einem Scheitern des Projekts führen.

- Es ist vorteilhaft, wenn die Erstellung der Dokumentation durch eine Person, die nicht Teil des Teams ist, begleitet wird. Ausufernde Diskussionen können so frühzeitig vermieden werden und der Fokus auf den zu bearbeitenden Aspekt wird geschärft. Dies trägt dazu bei, dass die Bearbeitung zu einem nutzbringenden Ergebnis führt.
- Die laufende Pflege und Weiterentwicklung der einzelnen Dokumente ist gerade in der Anfangsphase der Nutzung im Unternehmen wichtig, um gezielt auf Hürden eingehen zu können, durch die die Verwendung im Arbeitsalltag erschwert wird.



5 Diskussion und Ausblick

Um vertrauenswürdige KI im Entwicklungsprozess zu verankern und ein effektives KI-Qualitätsmanagement aufzubauen, sind Anpassungen an vielen Stellschrauben erforderlich. In dem vorliegenden Whitepaper wurde untersucht, mit welchen Methodiken ein bekanntes Entwicklungsparadigma (MLOps) insbesondere im Hinblick auf die Erzeugung hochwertiger Dokumentation angepasst werden kann, um Entwicklerinnen und Entwickler sowie Unternehmen dabei zu unterstützen. Dabei ist es notwendig, rechtliche und ethische Anforderungen von Beginn an in die KI-Entwicklung zu integrieren, um das Vertrauen von Stakeholdern und Nutzern zu gewinnen.

Neben der Erstellung übergreifender Dokumentationen wie dem KI-Steckbrief und dem Datasheets ist es insbesondere wichtig, die verschiedenen Lebensphasen eines KI-Systems – von der Entwicklung bis zum Betrieb – zu berücksichtigen und wichtige Entscheidungen strukturiert zu dokumentieren. Die Einführung und Etablierung von MLOps-Prinzipien als methodischer Rahmen, um KI-Projekte strukturiert, effizient und regelkonform umzusetzen, ist hierbei entscheidend. Das Lebensphasenmodell der KI-Entwicklung bildet die Grundlage für die Standardisierung von Prozessen. Dadurch wird die Grundlage geschaffen, um den Anforderungen des EU AI Acts und anderer potenzieller Audits gerecht zu werden und Qualität, Nachvollziehbarkeit und Compliance sicherzustellen.

Der erfolgreiche Einsatz von KI-Systemen geht weit über die reine Modellentwicklung hinaus. Erst durch die Verbindung von Monitoring, Governance, Standardisierung und praktischer Umsetzung im Unternehmenskontext entsteht ein Rahmen, der sowohl technische Exzellenz als auch regulatorische und ethische Anforderungen gleichermaßen adressiert.

Auf technischer Ebene wurde die Bedeutung der kontinuierlichen Überwachung und Bewertung der Modellleistung in den Mittelpunkt gestellt. Es wird deutlich, dass ohne ein systematisches Monitoring – etwa durch Leistungsmetriken, Drift- und Anomalie-Erkennung – die Zuverlässigkeit und Genauigkeit von KI-Modellen im produktiven Betrieb gefährdet ist. Die regelmäßige Aktualisierung der Modelle wird als entscheidender Schritt herausgearbeitet, um den sich verändernden Anforderungen und Datenumgebungen gerecht zu werden.

Am Beispiel des Unternehmens BITMARCK wurde anschaulich dargestellt, wie die Einführung der Methodiken ablaufen kann. Aus dem Unternehmenskontext heraus wurden gezielt

Anpassungen vorgenommen, um existierenden Prozessen und Anforderungen gerecht zu werden.

Abgesehen von klareren Formulierungen und angepassten Dokumenten für eine leichtere Verwendung des Leitfadens im Arbeitsalltag finden sich vor allem auf Seiten der Decision Records Möglichkeiten, den Dokumentationsaufwand für die Teams weiter zu reduzieren, ohne Abstriche bei der Einhaltung regulatorischer Anforderungen zu machen.

Einige Decision Records beschäftigen sich mit Themen, die von unternehmensweiter Tragweite sind. In diesem Fall bietet es sich an, Standardprozesse zu etablieren, die in den entsprechenden Lebensphasen umgesetzt werden oder bereitstehen und durch die Teams genutzt werden können. In diesem Fall werden Decision Records obsolet, da Entscheidungen nicht mehr individuell durch die Teams getroffen werden müssen, sondern bereits klar ist, welche Maßnahmen zu ergreifen sind. Ähnlich kann dies auch auf andere Komponenten des Leitfadens zutreffen. Standardisierte Integrationswege für ein Produkt können als Blaupause für weitere Produkte verwendet werden, die nach einem ähnlichen architektonischen Schema aufgebaut sind. Genauso gilt dies für Datenflüsse und -verarbeitungsschritte, die mehr und mehr standardisiert werden können. Dies trägt zu einer weiteren Entlastung einzelner Teams bei und kann die Entwicklung enorm beschleunigen.

Insgesamt hat sich der TAIOPs bereits jetzt als wirksamer Rahmen zur Sicherstellung der Entwicklung von vertrauenswürdigen und sicheren KI-Systemen etabliert. Durch die Verbindung regulatorischer Anforderungen mit praktischer Umsetzbarkeit schafft er eine solide Grundlage für ein nachhaltiges Qualitätsmanagement für die KI-Produktentwicklung.

Obwohl die hier vorgestellten Methodiken bereits wesentliche Anforderungen des EU AI Acts adressieren, muss stetig an einer weiteren Verbesserung der Prozesse gearbeitet werden. Die EU wird in den kommenden Monaten und Jahren harmonisierte Standards veröffentlichen, die darlegen, welche Tätigkeiten und Dokumentationsschritte zu erfolgen haben, um mit den gesetzlichen Anforderungen konform zu sein. Diese Entwicklungen müssen eng verfolgt werden, um auch auf Änderungen reagieren zu können und existierende Prozesse aktuell zu halten.

Autorinnen und Autoren



Lisa Fink arbeitet als KI-Technologiemanagerin bei der Kompetenzplattform KI.NRW. Sie beschäftigt sich mit dem Wissens- und Technologietransfer aktueller Entwicklungen im Bereich Künstliche Intelligenz und wirkt an Beratungs- und Forschungsprojekten zum vertrauenswürdigen Einsatz von KI mit. Zuvor studierte sie Informatik an der Hochschule Bonn Rhein Sieg.



Fabian Malms ist Projektleiter in der Abteilung AI Assurance and Assessments am Fraunhofer IAIS und leitet unter anderem das Flaggschiff-Projekt »Zertifizierte KI«. Er hat zwei Master of Laws (LL.M.) in Europäischem Recht und Wirtschaftsrecht von der Universität Maastricht. Der Schwerpunkt seiner Arbeit liegt auf KI-Governance und der Frage, wie durch ein KI-Qualitäts- und Risikomanagement eine »AI Act Compliance by design« in Organisationen erreicht werden kann.



Dr. Michael Mock ist Senior Data Scientist in der Abteilung AI Assurance and Assessments am Fraunhofer IAIS und Privat-Dozent für Informatik an der Universität Bonn. Mit über 120 wissenschaftlichen Publikationen sowie mehreren Software-Patenten hat er über 35 Jahren Erfahrung in der Forschung sowie in der Durchführung großer Kundenprojekte. Er war als wissenschaftlicher Koordinator von EU-weiten sowie nationalen Forschungsprojekten tätig und hat mehrere DAX-Konzerne sowie mittelständische Unternehmen in der Entwicklung und dem Einsatz KI-basierter Technologien beraten. Seine Forschungsgebiete umfassen die Absicherung und Prüfung von vertrauenswürdiger KI.



Dr. Maximilian Poretschkin leitet die Abteilung AI Assurance and Assessments am Fraunhofer IAIS und berät in dieser Funktion weltweit Unternehmen und Behörden zu vertrauenswürdiger KI. Seine Forschungsinteressen umfassen die informatische Operationalisierung rechtlicher Anforderungen, die Entwicklung von KI-Prüfmethoden, -Kriterien und -Werkzeugen sowie die Erarbeitung von KI-Governance Frameworks. Stationen vor Fraunhofer waren ein Postdoc an der University of Pennsylvania (Philadelphia, USA) und eine Tätigkeit als Consultant bei der Strategieberatung Bain & Company. Maximilian Poretschkin hat Physik und Mathematik in Bonn und Amsterdam studiert.



Lennard Helmer ist Senior Data Scientist in der Abteilung AI Assurance and Assessments am Fraunhofer IAIS. Er hat einen Master of Science in Business Analytics mit Schwerpunkt Mathematik und Informatik der Technischen Universität Freiberg und war mehrere Jahre in einem Beratungsunternehmen tätig. Sein Arbeits- und Forschungsschwerpunkt ist die Operationalisierung von Anforderungen von vertrauenswürdiger KI und KI-Regulierung im Entwicklungsprozess, welche er in verschiedenen Industrie- und Forschungsprojekten begleitet. Zudem ist er als Dozent am Data-Science-Schulungsprogramm beteiligt.



Claudio Martens ist im Team MLOps am Fraunhofer IAIS in der Beratung und Umsetzung sowie als Dozent für Schulungen im Bereich MLOps aktiv. Zu seinen Forschungsinteressen gehören Continuous Training, Model Monitoring und Drift Detection sowie Trustworthy AI Engineering und der Transfer dieser Themen in die Thematik LLMOps.



Dr. Benny Stein ist Leiter des Teams Machine Learning Operations am Fraunhofer IAIS. Er forscht an der Weiterentwicklung des MLOps-Zyklus für generative KI-Anwendungen und KI-Agenten, unter anderem zu verlässlichen Test-Strategien für derartige Systeme. Mit seinem Team führt er entlang des gesamten KI-Lebenszyklus Beratungs- und Implementierungsprojekte bei Kundinnen und Kunden durch. Zudem ist er als Dozent im Data-Science-Schulungsprogramm tätig.



Dr. Sermad Abbas ist Data Scientist im Center of Excellence KI bei BITMARCK und unterstützt Projektteams bei der Entwicklung datengetriebener Produkte unter Einsatz moderner MLOps-Praktiken. Seine Arbeitsschwerpunkte liegen in der konzeptionellen und operativen Weiterentwicklung von Entwicklungsprozessen für vertrauenswürdige KI-Lösungen. Besonders von Interesse sind für ihn hierbei Monitoring- und Evaluationsmethoden für KI-Systeme.

Literaturverzeichnis

Akbarighatar, P. »Operationalizing responsible AI principles through responsible AI capabilities.« *AI Ethics* 5, 2025: 1787–1801. <https://doi.org/10.1007/s43681-024-00524-4>.

Beck, Niklas, Claudio Martens, Karl-Heinz Sylla, Dennis Wegener, und Alexander Zimmermann. *Zukunftssichere Lösungen für maschinelles Lernen*. Sankt Augustin: Fraunhofer IAIS, 2020.

Bitkom e. V. »Generative KI im Unternehmen.«, 2025; abrufbar unter: <https://www.bitkom.org/sites/main/files/2024-02/Bitkom-Leitfaden-Generative-KI-im-Unternehmen.pdf> (zuletzt aufgerufen am 28. Januar 2026).

Gebri, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iijima, H. D., & Crawford, K. »Datasheets for datasets.« *Communications of the ACM*, 64(12), 2021: 86-92.

Gillespie, Nicole, Steve Lockey, Tabi Ward, Alexandria Macdade, und Gerard Hased. »Trust, Attitudes and Use of Artificial Intelligence: A Global Study 2025.« The University of Melbourne and KPMG, 2025.

Helmer, Lennard, Claudio Martens, Dennis Wegener, Maram Akila, Daniel Becker, und Sermad Abbas. »Towards Trustworthy AI Engineering-A Case Study on integrating an AI audit catalog into MLOps processes.« *Proceedings of the 2nd International Workshop on Responsible AI Engineering*. New York, NY, USA: Association for Computing Machinery, 2024.

Kreuzberger, Dominik, Niklas Kühl, und Sebastian Hirschl. »Machine Learning Operations (MLOps): Overview, Definition, and Architecture.« *IEEE Access* (Vol. 11), 2023: 31866–31879.

Mäntymäki, Matti, Matti Minkkinen, Teemu Birkstedt, und Mika Viljanen. »Defining Organizational AI Governance.« *AI and Ethics* (Springer) 2, Nr. 4, 2022: 603-609.

Mittelstadt, Brent. »Principles Alone Cannot Guarantee Ethical AI.« *Nature Machine Intelligence* (Nature Publishing Group UK London) 1, Nr. 11, 2019: 501-507.

Poretschkin, Maximilian et al. *Leitfaden zur Gestaltung vertrauenswürdiger Künstlicher Intelligenz (KI-Prüfkatalog)*. Sankt Augustin: Fraunhofer IAIS, 2021.

Schmitz, Anna, Maram Akila, Dirk Hecker, Maximilian Poretschkin, und Stefan Wrobel. »The Why and How of Trustworthy AI.« *at - Automatisierungstechnik* (De Gruyter (O)) 70, Nr. 2, 2022: 793-804.

Impressum

Herausgeber

Fraunhofer-Institut für Intelligente Analyse-
und Informationssysteme IAIS
Schloss Birlinghoven 1
53757 Sankt Augustin

Redaktion

Julia Kabbalo

Grafik und Layout

Achim Kapusta
Asra-Soraya Neumeister

Titelbild

Alex - stock.adobe.com

Stand

Februar 2026



Kontakt

Fraunhofer-Institut für Intelligente
Analyse- und Informationssysteme IAIS
Schloss Birlinghoven 1
53757 Sankt Augustin

www.iais.fraunhofer.de

Ansprechpartner:
Lennard Helmer
lennard.helmer@iais.fraunhofer.de