Fraunhofer IAIS

Leitfaden zur Gestaltung vertrauenswürdiger Künstlicher Intelligenz

KI-Prüfkatalog

POWERED BY KI NRW

Gefördert durch:

Ministerium für Wirtschaft, Innovation, Digitalisierung und Energie des Landes Nordrhein-Westfalen

ZERTIFIZIERTE KI
Qualität sichern. Fortschritt gestalten.

Fraunhofer IAIS

Leitfaden zur Gestaltung vertrauenswürdiger Künstlicher Intelligenz

KI-Prüfkatalog

Autorinnen und Autoren

Dr. Maximilian Poretschkin
Anna Schmitz
Dr. Maram Akila
Linara Adilova
Dr. Daniel Becker
Prof. Dr. Armin B. Cremers
Dr. Dirk Hecker
Dr. Sebastian Houben
PD Dr. Michael Mock
Julia Rosenzweig
Joachim Sicking
Elena Schulz
Dr. Angelika Voss
Prof. Dr. Stefan Wrobel

www.iais.fraunhofer.de/ki-pruefkatalog

In Kooperation mit

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

ZERTIFIZIERTE KI
Qualität sichern. Fortschritt gestalten.

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

Inhalt

Grußwort

Liebe Leserinnen, liebe Leser,

Künstliche Intelligenz ist die Schlüsseltechnologie unserer Gegenwart. Intelligente Systeme finden sich mittlerweile in nahezu allen gesellschaftlichen Lebensbereichen wieder und unterstützen den Menschen dabei, Aufgaben schneller und zuverlässiger zu bewältigen. Prognosen lassen erwarten, dass Künstliche Intelligenz das weltweite Wirtschaftswachstum signifikant erhöhen und die Entwicklung der Gesellschaft entscheidend prägen wird – auch weil KI das Potenzial birgt, einen wichtigen Beitrag zur Bewältigung großer gesellschaftlicher Herausforderungen in Bereichen wie Klimaschutz, Mobilität und Gesundheit zu leisten.

In Deutschland verfügen wir über eine exzellente Ausgangslage im Bereich der Künstlichen Intelligenz und des Maschinellen Lernens – insbesondere in Nordrhein-Westfalen. Unser Industrie- und Innovationsland kann auf eine langjährige Erfahrung im Bereich der Künstlichen Intelligenz zurückgreifen und nimmt eine Vorbildfunktion ein, wenn es darauf ankommt, Wirtschaft und Forschung zu verzahnen und anwendungsorientierte KI-Forschung voranzutreiben. Bereits 2019 wurde zu diesem Zweck die Kompetenzplattform KI.NRW initiiert, um die Akteure im Bereich der Künstlichen Intelligenz zu vernetzen und den Technologietransfer von der Forschung in die Praxis zu stärken.

Mit dem Einsatz von KI-Anwendungen gehen allerdings auch spezifische Risiken einher, die Themen wie Sicherheit, Transparenz, Verlässlichkeit, Fairness, Autonomie und Datenschutz betreffen. Wenn wir das immense Potenzial der Technologie erschließen wollen, ist es unerlässlich, den Menschen in den Mittelpunkt der Gestaltung von KI-Anwendungen zu stellen, die erwähnten Risiken zu evaluieren und entsprechend zu minimieren. Eine perspektivisch von unabhängigen Prüfstellen ausgestellte Zertifizierung von KI-Produkten kann helfen, die Qualität der Systeme zu erhöhen, das Vertrauen und die Akzeptanz von KI in der Gesellschaft zu stärken und vor allem die Wettbewerbsfähigkeit unserer Unternehmen nachhaltig und über die Landesgrenzen hinaus zu sichern.

Der Ihnen nun vorliegende Prüfkatalog des Fraunhofer IAIS ist ein wichtiger Meilenstein auf dem Weg zu einer unabhängigen KI-Prüfung und beschreibt auf über 160 Seiten, wie KI-Anwendungen systematisch hinsichtlich Risiken evaluiert werden können, formuliert Vorschläge für Prüfkriterien zur Messung der Qualität der Systeme und schlägt Maßnahmen vor, die KI-Risiken mindern können. Damit dient der Prüfkatalog nicht nur als praxistauglicher Leitfaden, um gemäß einer vereinheitlichten Vorgehensweise produktspezifische, reproduzierbare und standardisierte Prüfverfahren und Qualitätsverbesserungen von KI-Systemen zu ermöglichen, sondern auch als nationaler und internationaler Weichensteller für eine ebenso innovationsfreundliche wie vertrauenswürdige Normierung und Standardisierung »made in Germany«. Die Wissenschaftlerinnen und Wissenschaftler aus Sankt Augustin haben den Prüfkatalog bereits in ersten Pilotprüfungen mit Unternehmen erfolgreich getestet.

Ich freue mich, dass wir das Vorhaben im neuen KI.NRW-Flagship »ZERTIFIZIERTE KI« gemeinsam weiter vorantreiben können und wünsche allen Beteiligten viel Erfolg.

Mit herzlichen Grüßen

Prof. Dr. Andreas Pinkwart
Minister für Wirtschaft, Innovation, Digitalisierung und Energie des Landes Nordrhein-Westfalen

Executive Summary

Künstliche Intelligenz (KI) hat in den vergangenen Jahren beeindruckende Fortschritte erzielt und prägt als Schlüsseltechnologie Wirtschaft und Gesellschaft entscheidend. Prominente Anwendungsbeispiele finden sich in der medizinischen Diagnostik, der prädiktiven Wartung und perspektivisch beim autonomen Fahren. Gleichzeitig liegt es auf der Hand, dass KI und auf ihr basierende Geschäftsmodelle nur dann ihr volles Potenzial entfalten können, wenn KI-Anwendungen nach hohen Qualitätsstandards entwickelt werden und wirksam gegen neuartige KI-Risiken abgesichert sind. Ein Beispiel eines solchen KI-Risikos ist etwa die ungerechtfertigte Diskriminierung von Nutzer*innen bei der KI-basierten Verarbeitung von personenbezogenen Daten zur Kreditvergabe oder Personalauswahl. Ein anderes Beispiel stellen schwerwiegende Falschprädiktionen dar, welche aus geringen Störungen in den Eingabedaten resultieren, etwa wenn Fußgänger*innen durch ein autonomes Fahrzeug aufgrund verrauschter Bildaufnahmen nicht erkannt werden. Dass diese neuen Risiken auftreten, ist eng verbunden mit der Tatsache, dass sich der Entwicklungsprozess von KI-Anwendungen, insbesondere solchen, die auf Maschinellem Lernen (ML) basieren, stark von dem herkömmlicher Software unterscheidet. Denn wie sich KI-Anwendungen verhalten, wird im Wesentlichen aus großen Datenmengen erlernt und nicht durch die Programmierung fester Regeln vorgegeben.

Die Frage nach der Vertrauenswürdigkeit von KI-Anwendungen ist daher zentral und Gegenstand vieler wichtiger Veröffentlichungen von Stakeholdern aus Politik, Wirtschaft und Gesellschaft. Zu nennen sind hier insbesondere der Verordnungsentwurf der Europäischen Kommission[1], die Normungsroadmap KI[2] und die Empfehlungen der High-Level Expert Group on AI[3], welche wichtige Leitplanken zum vertrauenswürdigen Einsatz von Künstlicher Intelligenz formulieren. Gleichzeitig besteht Einigkeit darin, dass in einem nächsten Schritt die oftmals abstrakt beschriebenen Anforderungen an vertrauenswürdige KI konkretisiert und greifbar gemacht werden müssen. Eine Herausforderung dabei ist, dass die konkreten Qualitätskriterien für eine KI-Anwendung stark vom Anwendungskontext und mögliche Maßnahmen zu deren Erfüllung wiederum stark von der verwendeten KI-Technologie abhängen. So sind bei einer automatisierten Analyse von Bewerbungsdokumenten die Anforderungen an die Vertrauenswürdigkeit eines KI-Systems anders zu bewerten als beispielsweise bei einem Bilderkennungsverfahren zur Qualitätssicherung von Autokarosserien. Schließlich werden praxistaugliche Prüfverfahren benötigt, um für spezifische KI-Anwendungen beurteilen zu können, ob diese nach angemessenen Qualitätsstandards entwickelt wurden.

Der vorliegende KI-Prüfkatalog setzt genau an dieser Stelle an und richtet sich an zwei Zielgruppen: Zum einen gibt er Entwickler*innen eine Richtschnur an die Hand, um ihre KI-Anwendungen systematisch vertrauenswürdig zu gestalten. Zum anderen leitet er Prüfer*innen dazu an, KI-Anwendungen strukturiert auf Vertrauenswürdigkeit zu untersuchen.

---

1 European Commission, Directorate-General for Communications Networks, Content and Technology (April 2021). Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. COM/2021/206 final https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206 (letzter Aufruf: 23.06.2021)

2 Wahlster, Winterhalter (Hrsg.) (November 2020). Deutsche Normungsroadmap Künstliche Intelligenz. Deutsches Institut für Normung e.V. und Deutsche Kommission Elektrotechnik. https://www.dke.de/de/arbeitsfelder/core-safety/normungsroadmap-ki (letzter Aufruf: 23.06.2021)

3 High-Level Expert Group on AI (HLEG) (April 2019). Ethics Guidelines on trustworthy AI. Europäische Kommission. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (letzter Aufruf: 21.06.2021)

Hierzu formuliert er ein vierstufiges Vorgehen:

1. Eine umfassende Risikoanalyse entlang der Dimensionen Fairness, Autonomie und Kontrolle, Transparenz, Verlässlichkeit, Sicherheit und Datenschutz.

2. Die Festlegung objektiver, möglichst messbarer Zielvorgaben, um die Mitigation der unter 1 identifizierten Risiken nachweisbar zu machen.

3. Eine systematische Auflistung von Maßnahmen entlang des Lebenszyklus einer KI-Anwendung, um die in 2 gesetzten Zielvorgaben zu erreichen.

4. Die Erstellung einer stringenten Argumentation, dass die unter 2 formulierten Zielvorgaben erreicht wurden (»Absicherungsargumentation für die Vertrauenswürdigkeit«), wobei auch KI-spezifische Trade-Offs, z. B. Sicherheit vs. Transparenz, berücksichtigt werden.

1. Einleitung

Künstliche Intelligenz (KI) dringt in immer mehr Bereiche unseres Alltagslebens vor und übernimmt zunehmend verantwortungsvolle Aufgaben. Beispielhaft zu nennen sind die KI-basierte Qualitätskontrolle in der Produktion, Unterstützungssysteme für die medizinische Diagnostik, automatisierte Börsengeschäfte oder perspektivisch das autonome Fahren. Für KI-basierte Geschäftsmodelle ist es hierbei von zentraler Bedeutung, dass die KI-Anwendung verlässlich, sicher und resilient ist. Gleichzeitig ist es für den Menschen als Nutzenden und Betroffenen wichtig, dass der Einsatz von KI im Einklang mit gesellschaftlichen Wertvorstellungen erfolgt. Es zeigt sich somit, dass insbesondere für sensible Anwendungskontexte das Potenzial von KI nur dann voll ausgeschöpft werden kann, wenn KI-Anwendungen nach hohen Qualitätsmaßstäben realisiert werden.

Dabei stellen sich verschiedene Herausforderungen einerseits in Bezug auf die Verantwortung für die Qualität von KI-Anwendungen und andererseits hinsichtlich der technischen Überprüfbarkeit von Qualitätsanforderungen.

Die Verantwortung für die Qualität von KI-Systemen ist entlang einer Wertschöpfungskette verteilt, die sich stark von der Entwicklung herkömmlicher Software unterscheidet. KI-Anwendungen basieren oftmals auf Verfahren des Maschinellen Lernens (ML), die Muster in sogenannten Trainingsdaten lernen und ein Modell erstellen, um das Gelernte auf unbekannte (aber strukturell mit den Trainingsdaten vergleichbare) Daten anzuwenden. Da derartige Modelle oft über Millionen (bisweilen auch Milliarden) an Parametern spezifiziert werden, beruhen KI-Anwendungen somit insbesondere auf der Verarbeitung großer Datenmengen, wofür entsprechende IT-Infrastrukturen und Rechenleistungen benötigt werden. Neben den Datenerzeugern spielen hierbei auch Cloud-Dienstleister eine wichtige Rolle, welche die benötigte Rechenkapazität, Infrastruktur sowie entsprechende KI-Basisdienstleistungen wie beispielsweise Optical Character Recognition (OCR) bereitstellen und somit maßgeblich die Qualität von KI-Anwendungen mitbeeinflussen können.

Neben einer komplexen Wertschöpfungskette stellt auch die Komplexität der KI-Anwendungen selbst eine Herausforderung für die Sicherstellung ihrer Qualität dar. So kann die Funktionsweise der zugrundeliegenden Modelle, etwa aufgrund der hohen Anzahl an Parametern, selbst von Expert*innen oftmals nur schwer nachvollzogen werden. Des Weiteren kann sich der Lernprozess prinzipiell auch während des Betriebs der KI-Anwendungen fortsetzen, sodass Leitplanken benötigt werden, um beispielsweise das Erlernen von Fehlverhalten zu vermeiden.

Prüfung als wichtiger Baustein für Vertrauen und Qualität

Es ist daher für KI-Expert*innen wichtig, Qualität systematisch in die Entwicklung eigener KI-Anwendungen implementieren und gleichzeitig die Qualität dritter Systeme einschätzen zu können. Ferner müssen Nutzer*innen und Betroffene darauf vertrauen können, dass entsprechende KI-Anwendungen angemessenen Qualitätsanforderungen genügen. Ein in anderen Domänen bewährter Baustein, um Vertrauen herzustellen, sind sachkundige und neutrale Prüfungen. Marktfähige Prüfverfahren, die zugesicherte Eigenschaften von KI-Produkten und -Dienstleistungen bestätigen, können beispielsweise einen Beitrag zur Markenbildung leisten und somit Wettbewerbsvorteile schaffen. Daneben können Prüfungen auch Teil von vorgeschriebenen Zulassungs- und Aufsichtsverfahren sein. Mit dem kürzlich veröffentlichten Verordnungsentwurf der Europäischen Kommission[4]

4 European Commission, Directorate-General for Communications Networks, Content and Technology (April 2021). Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. COM/2021/206 final https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206 (letzter Aufruf: 23.06.2021)

zeichnet sich ab, dass solche Zulassungsverfahren im Bereich KI für den europäischen Markt zeitnah etabliert werden. Denn neben dem Verbot gewisser Einsatzzwecke für KI gibt der Verordnungsentwurf vor, dass »Hochrisikosysteme« einer Konformitätsbewertung unterzogen werden müssen. Dies betrifft eine Vielzahl von KI-Anwendungen, die bereits fester Bestandteil unseres Alltagslebens sind.

Operationalisierung von Qualitätsanforderungen und bestehenden KI-Richtlinien

Während allgemeine Anforderungen an die Vertrauenswürdigkeit von KI bereits seit geraumer Zeit Gegenstand intensiver gesellschaftlicher und politischer Diskussionen sind und viele Richtlinien zur Vertrauenswürdigkeit von KI-Anwendungen veröffentlicht wurden, ist die Operationalisierung dieser Richtlinien noch immer zu großen Teilen offen. Deutlich wird dies insbesondere daran, dass die systembezogenen Anforderungen im Verordnungsentwurf der Europäischen Kommission nicht als eindeutige bzw. quantitative Kriterien formuliert sind, sondern großen Ermessensspielraum lassen. Für die technische Konkretisierung der Anforderungen verweist die Europäische Kommission auf harmonisierte Normen und technische Standards[5], die jedoch, wie die im Dezember 2020 veröffentlichte Normungsroadmap KI[6] überdeutlich aufzeigt, in großen Teilen noch nicht vorhanden sind. Auch das Bundesamt für Sicherheit in der Informationstechnik (BSI) hat den Handlungsbedarf bezüglich der Entwicklung von »Standards, technischen Richtlinien, Prüfkriterien und Prüfmethoden«[7] für den sicheren Einsatz von KI-Anwendungen betont.

Festzuhalten ist hierbei, dass die konkreten Anforderungen, die zur Erfüllung der Vertrauenswürdigkeit an eine KI-Anwendung zu stellen sind, hochgradig von der verwendeten Technologie und dem Einsatzkontext abhängen. Wünschenswert wären Key Performance Indicators (KPIs), die die Qualität von KI-Anwendungen messbar machen. Die Normungsroadmap KI diskutiert hierfür das Beispiel von KI-basierten Übersetzungssystemen, wobei die Qualität der Übersetzung durch den sogenannten BLEU Score[8] gemessen wird. Der BLEU Score vergleicht hierbei auf einer Skala von 0 bis 100 eine KI-basierte Übersetzung mit einer menschlichen Übersetzung, wobei der Wert 100 die perfekte Übereinstimmung bedeutet. Je kritischer der Verwendungszweck der Übersetzung, desto höher muss der BLEU Score sein. In dem Beispiel wird etwa für die Übersetzung von Social Media Posts ein BLEU Wert von mindestens 35 vorgeschlagen und für die Übersetzung von Arztbriefen ein BLEU Wert von mindestens 45. Angesichts der Fülle von KI-Technologien und ihrer unterschiedlichen Anwendungskontexte stellt sich jedoch die Frage, welche Metrik und welcher Schwellwert für einen gegebenen Anwendungskontext verhältnismäßig sind. Im Beispiel der Vermeidung einer ungerechtfertigten Diskriminierung durch eine KI-Anwendung wird aus technischer Sicht eine Quantifizierung von Fairness benötigt. Zur Fairnessquantifizierung stehen jedoch eine Vielzahl unterschiedlicher Konzepte und Metriken zur Verfügung. Somit muss im Einzelfall entschieden werden, welche Kriterien und Schwellwerte angemessen sind, um die Fairness einer KI-Anwendung zu beurteilen.

Bei der Festlegung konkreter Qualitätsanforderungen für KI-Anwendungen besteht zusätzlich die Herausforderung, dass die unterschiedlichen Dimensionen der Vertrauenswürdigkeit nicht unabhängig voneinander beurteilt werden können, sondern Zielkonflikte möglich sind. So kann eine Erhöhung der Performanz, wie zum Beispiel

---

5 »The precise technical solutions to achieve compliance with those requirements may be provided by standards or by other technical specifications or otherwise be developed in accordance with general engineering or scientific knowledge at the discretion of the provider of the AI system.« (S. 13) und »common normative standards for all high-risk AI systems should be established« (S. 20):
European Commission, Directorate-General for Communications Networks, Content and Technology (April 2021). Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. COM/2021/206 final https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206 (letzter Aufruf: 23.06.2021)

6 Wahlster, Winterhalter (Hrsg.) (November 2020). Deutsche Normungsroadmap Künstliche Intelligenz. Deutsches Institut für Normung e. V. und Deutsche Kommission Elektrotechnik. https://www.dke.de/de/arbeitsfelder/core-safety/normungsroadmap-ki (letzter Aufruf: 23.06.2021)

7 Bundesamt für Sicherheit in der Informationstechnik (Februar 2021). Sicherer, robuster und nachvollziehbarer Einsatz von KI. https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/KI/Herausforderungen_und_Massnahmen_KI.pdf?__blob=publicationFile&v=5 (letzter Aufruf: 23.06.2021)

8 BLEU steht für Bilingual Evaluation Understudy

der Erkennungsleistung bei der Objekterkennung auf Bilddaten durch sog. tiefe Neuronale Netze, zu Lasten der Nachvollziehbarkeit gehen oder eine Erhöhung der Transparenz (etwa durch die Offenlegung aller Hyperparameter eines Modells) zu neuen Angriffsvektoren im Sinne der IT-Sicherheit führen.

Risikobasierte KI-Prüfung

Eine bewährte Herangehensweise, mit der sich anwendungsbezogen spezifische Qualitätsanforderungen festlegen lassen, ist ein risikobasierter Prüfansatz. Risikobasierte Prüfansätze haben sich bereits in den Bereichen der klassischen IT-Sicherheit[9] und Funktionalen Sicherheit bewährt, wo die Forderung nach der Resistenz gegenüber Manipulationen oder ungewolltem Fehlverhalten für unterschiedliche IT-Systeme zu sehr unterschiedlichen technischen Anforderungen führen kann. Hierbei ist es möglich, Risiken unter verschiedenen Gesichtspunkten zu betrachten: In erster Linie wird das Risiko einer Fehlfunktion hinsichtlich der Auswirkungen auf Nutzer*innen, Betroffene oder die (unmittelbare) Umgebung untersucht. Dabei werden sowohl potenzielle materielle als auch immaterielle Schäden betrachtet, beispielsweise in Bezug auf Sicherheit und Persönlichkeitsrechte. Gleichzeitig sind mit fehlerhaftem oder gar schädlichem Verhalten einer KI-Anwendung auch immer Risiken für Organisationen bzw. die Personen verbunden, welche die Verantwortung für die KI-Anwendung tragen. So besteht bei einer KI-Anwendung zur Kreditvergabe das Risiko der Diskriminierung, womit einerseits ein immaterieller Schaden für Betroffene verbunden ist, andererseits eine Rufschädigung des betreffenden Kreditinstituts. Dies zeigt, dass sich die Diskussion der vertrauenswürdigen KI nicht allein auf systemspezifische Fragestellungen einer KI-Anwendung beziehen kann, sondern in die Entwicklung einer KI-Governance und somit einer entsprechenden Unternehmenskultur übergehen muss[10].

Ein risikobasierter Prüfansatz für KI bietet darüber hinaus die Möglichkeit, dass er sich gut in bestehende Prüfschemata integrieren lässt. Die Anschlussfähigkeit an bestehende Prüfschemata und -verfahren ist für die Praktikabilität und Marktfähigkeit eines KI-Prüfverfahrens von zentraler Bedeutung, da KI-Prüfungen in vielen Bereichen als ein Teil einer übergeordneten Prüffragestellung durchgeführt werden, die auch andere, nicht KI-spezifische Fragestellungen zur Software und Hardware adressiert. So fordert der Verordnungsentwurf der Europäischen Kommission explizit, dass die Konformitätsbewertung gewisser KI-basierter Hochrisikosysteme, abhängig vom Anwendungsbereich, etwa in die bestehende Prüfprozedur zur Produktsicherheit oder in die etablierten Prozesse zur aufsichtlichen Überprüfung von Kreditinstituten integriert werden soll[11].

Leistungsspektrum und Anwendungsgebiete des KI-Prüfkatalogs

Der vorliegende Prüfkatalog bietet einen strukturierten Leitfaden, anhand dessen KI-Anwendungen in Bezug auf alle relevanten Dimensionen der Vertrauenswürdigkeit evaluiert werden können. Hierbei fokussiert er auf KI-Anwendungen, die auf Maschinellem Lernen beruhen. Der Katalog formuliert dabei einen primär produktorientierten Prüfansatz. Da sich KI-Anwendungen jedoch während des Betriebs dynamisch weiterentwickeln können, formuliert er zusätzlich organisatorisch-prozessuale Vorgaben, die essenziell sind, um die Vertrauenswürdigkeit der KI-Anwendung auch nach dem Zeitpunkt der Beurteilung sicherzustellen.

---

9 Siehe etwa den BSI-Grundschutz oder die ISO 27001 (ISO/IEC 27001, 2013) mit Vorgaben für ein IT-Sicherheitsmanagementsystem oder die Common Criteria (CC 3.1, 2017) für eine Methodik zur Prüfung von IT-Produkten.

10 Für eine ausführliche Betrachtung siehe die Fraunhofer IAIS Studie »AI Management Systems« (in Erscheinung), die Anforderungen an Organisationen zum Umgang mit KI in Bezug auf Governance, Management und technisch-organisatorische Maßnahmen diskutiert und dabei die aktuellen Standardisierungsaktivitäten des ISO/IEC JTC1/SC 42 »Artificial Intelligence« miteinbezieht.

11 »The key difference is that the ex-ante and ex-post mechanisms will ensure compliance not only with the requirements established by sectorial legislation, but also with the requirements established by this regulation.« (S. 13) und siehe auch Artikel 43 in: European Commission, Directorate-General for Communications Networks, Content and Technology (April 2021). Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. COM/2021/206 final https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206 (letzter Aufruf: 23.06.2021)

Insbesondere bietet der Katalog:

1. Einen Leitfaden zur strukturierten Identifikation KI-spezifischer Risiken in Hinblick auf sechs Dimensionen der Vertrauenswürdigkeit: Fairness, Autonomie und Kontrolle, Transparenz, Verlässlichkeit, Sicherheit und Datenschutz[12].

2. Eine Anleitung, mit deren Hilfe spezifische Prüfkriterien für eine KI-Anwendung formuliert werden können. Zu diesem Zweck listet der KI-Prüfkatalog etablierte KPIs (vergleichbar zu dem oben erwähnten BLEU Score) auf, mit denen sich entsprechende Zielvorgaben, wo möglich, quantifizieren lassen.

3. Eine Anleitung zur strukturierten Dokumentation von technischen und organisatorischen Maßnahmen entlang des Lebenszyklus einer KI-Anwendung, die dem aktuellen Stand der Technik entsprechen und durch deren Umsetzung mögliche Risiken abgeschwächt werden können.

Einordnung in bestehende Prüfansätze

Der hier vorgestellte KI-Prüfkatalog ergänzt bestehende Ansätze zur Prüfung von KI-Anwendungen, unter denen als besonders prominent der AIC4-Katalog des BSI[13] sowie das Framework zur Operationalisierung von KI-Ethik der AI Ethics Impact Group[14] zu nennen sind:

Der BSI-Katalog stellt eine Erweiterung des C5-Katalogs[15] dar und kann zur Prüfung von Cloud-basierten KI-Diensten verwendet werden. Die darin enthaltenen Kriterien decken den gesamten Lebenszyklus eines KI-Dienstes ab und adressieren die Bereiche Robustheit, Performanz, Zuverlässigkeit, Datenqualität, Erklärbarkeit und Bias. Dabei werden jedoch weder die Anforderungen in Bezug auf diese KI-spezifischen Risiken konkretisiert noch Hinweise gegeben, wie die Erfüllung dieser Anforderungen im konkreten System zu bewerten ist.

Das Framework zur Operationalisierung von KI-Ethik der AI Ethics Impact Group zielt darauf ab, die Konformität einer KI-Anwendung mit ethischen Prinzipien wie »Transparenz, Rechenschaftspflicht, Privatheit, Gerechtigkeit, Verlässlichkeit und Nachhaltigkeit« messbar zu machen. Hierfür wird ein Wert-Analyseverfahren aus einer Kombination von Zielkriterien, Indikatoren und messbaren Größen verwendet, um Qualitätsstufen analog zu den Energieeffizienzklassen für Elektrogeräte abzuleiten. Dieser Ansatz fokussiert darauf, wichtige Qualitätseigenschaften für den*die Endverbraucher*in transparent darzustellen, verzichtet aber auf eine tiefergehende informatische Untersuchung der KI-Anwendung.

Der KI-Prüfkatalog ergänzt die bestehenden Ansätze, indem er ein systematisches Vorgehen zur Entwicklung einer Absicherungsargumentation für KI-Anwendungen auf einer konkreteren Ebene aufzeigt. Die dargestellten

---

12 Die sechs Dimensionen der Vertrauenswürdigkeit sind angelehnt an die Kernanforderungen für vertrauenswürdige KI der High-Level Expert Group on AI, siehe dazu: High-Level Expert Group on AI (HLEG) (April 2019). Ethics Guidelines on Trustworthy AI. Europäische Kommission. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (letzter Aufruf: 21.06.2021). Eine ausführliche Beschreibung der sechs im Prüfkatalog betrachteten Dimensionen ist zu finden in: Poretschkin, M., et al. (2019). Vertrauenswürdiger Einsatz von Künstlicher Intelligenz. Sankt Augustin: Fraunhofer-Institut für Intelligente Analyse und Informationssysteme IAIS. https://www.ki.nrw/wp-content/uploads/2020/03/Whitepaper_KI-Zertifizierung.pdf (letzter Aufruf: 18.06.2021)

13 Bundesamt für Sicherheit in der Informationstechnik (BSI) (Februar 2021). AI Cloud Service Compliance Criteria Catalogue (AIC4). Bundesamt für Sicherheit in der Informationstechnik (BSI). https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/CloudComputing/AIC4/AI-Cloud-Service-Compliance-Criteria-Catalogue_AIC4.pdf?__blob=publicationFile&v=4 (letzter Aufruf: 23.06.2021)

14 Hallensleben, S. und Hustedt, C. et al. (2020). From Principles to Practice, An interdisciplinary framework to operationalise AI ethics. Bertelsmann Stiftung. https://www.ai-ethics-impact.org/resource/blob/1961130/c6db9894ee73aefa489d6249f5ee2b9f/aieig---report---download-hb-data.pdf (letzter Aufruf: 23.06.2021)

15 Bundesamt für Sicherheit in der Informationstechnik (BSI) (Januar 2020). Cloud Computing Compliance Criteria Catalogue – C5:2020. Kriterienkatalog Cloud Computing. Bundesamt für Sicherheit in der Informationstechnik (BSI). https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/CloudComputing/Anforderungskatalog/2020/C5_2020.pdf?__blob=publicationFile&v=2 (letzter Aufruf: 23.06.2021)

risikoabschwächenden Vorkehrungen und Testmaßnahmen bieten insbesondere für Entwickler*innen bzw. Betreiber*innen von Hochrisikosystemen eine Hilfestellung, die gemäß dem Regulierungsentwurf der Europäischen Kommission zukünftig einen Nachweis über die Erfüllung technischer Qualitätsanforderungen erbringen müssen. Hierzu fordert die Europäische Kommission, dass Betreiber*innen eine technische Dokumentation der KI-Anwendung vorlegen, auf deren Grundlage eine Konformitätsbewertung, ggf. durch behördliche Prüfstellen, durchgeführt wird. Der vorliegende Prüfkatalog bietet nun einen Leitfaden zur Anfertigung einer solchen technischen Dokumentation[16]. Dabei deckt er unter anderem die Beschreibung der wesentlichen Entwicklungsschritte, die Abschätzung der KI-spezifischen Risiken, die Formulierung von Kriterien bzw. Metriken zur Quantifizierung relevanter technischer Eigenschaften, die Spezifikation der ergriffenen risikoabschwächenden Maßnahmen über den Lebenszyklus hinweg sowie schließlich die Diskussion möglicher Trade-Offs ab.

Zusammenfassend können unabhängige Prüfer*innen die anhand des KI-Prüfkatalogs angefertigte Dokumentation als Grundlage zur Beurteilung der KI-Anwendung sowie zur Planung tiefergehender Untersuchungen bzw. Tests verwenden. Außerdem bietet der Prüfkatalog eine Richtschnur für Entwickler*innen, mit deren Hilfe sie vertrauenswürdige KI-Anwendungen nach dem aktuellen Stand der Technik gestalten können.

Aufbau des Katalogs

In Kapitel 2 wird erläutert, wie der KI-Prüfkatalog konkret zur Beurteilung der Vertrauenswürdigkeit von KI-Anwendungen eingesetzt werden kann. Dazu werden zunächst Grundlagen und wichtige Begriffe eingeführt und anschließend die risikobasierte Systematik des Prüfkatalogs im Detail erläutert. Nach dieser Systematik wird die Vertrauenswürdigkeit von KI-Anwendungen hinsichtlich der sechs Dimensionen Fairness, Autonomie und Kontrolle, Transparenz, Verlässlichkeit, Sicherheit sowie Datenschutz beurteilt. Im dritten Kapitel wird der KI-Steckbrief vorgestellt, mit dessen Hilfe ein Überblick über die KI-Anwendung gegeben und der Prüfgegenstand eingegrenzt wird. Die nachfolgenden Kapitel bieten einen Leitfaden, anhand dessen KI-Risiken in Hinblick auf die sechs Dimensionen der Vertrauenswürdigkeit strukturiert bewertet werden können. Im letzten Kapitel wird das Vorgehen zur abschließenden dimensionsübergreifenden Beurteilung beschrieben.

16 Die Anforderungen der Europäischen Kommission an die technische Dokumentation werden in Annex IV des Verordnungsentwurfs beschrieben. Der KI-Prüfkatalog deckt die in Punkt 1, 2, 3, 5 und z.T. in 8 von Annex IV geforderten Inhalte ab.

2. Grundlegende Konzepte und Methodik zur Anwendung des Katalogs

Die Qualitätsanforderungen an eine KI-Anwendung ergeben sich maßgeblich aus ihrem Einsatzzweck und -kontext. Wie diese Qualitätsanforderungen umgesetzt werden, um eine vertrauenswürdige KI-Anwendung zu gestalten, hängt zudem stark von der zugrundeliegenden KI-Technologie ab.

Der KI-Prüfkatalog bietet ein Framework, mit dessen Hilfe die Beurteilung der Vertrauenswürdigkeit strukturiert durchgeführt werden kann. Eine wichtige Voraussetzung zur Anwendung des Katalogs ist ein klar definierter Prüfgegenstand[17], welcher ggf. von einem umgebenden Gesamtsystem abgegrenzt werden muss. Daher werden in diesem Kapitel zunächst inhaltliche und konzeptuelle Grundlagen hinsichtlich des formalen Aufbaus einer KI-Anwendung vermittelt (Abschnitt 2.1). Ferner wird erläutert, welche Risiken die Qualitätskriterien des KI-Prüfkatalogs inhaltlich abdecken (Abschnitt 2.2) und wie sich eine KI-Prüfung entlang dieses Katalogs konkret gestalten kann[18] (Abschnitt 2.3).

KI-Anwendungen sind komplexe Konstrukte, innerhalb derer implementierte ML-Modelle meist mit Expertensystemen und anderen klassischen Softwarekomponenten zusammenspielen; oder die gar als hybride Systeme entworfen sind, die etwa strukturiertes Wissen z. B. Wissensgraphen (Knowledge Graphs), mit Maschinellem Lernen kombinieren. Zudem sind KI-Anwendungen oftmals in ein größeres Gesamtsystem eingebettet. Die Frage, wie sich einzelne Komponenten sowie die KI-Anwendung innerhalb eines Gesamtsystems abgrenzen, ist in der Literatur nicht durch eindeutige Begrifflichkeiten geregelt. Auch ist zunächst offen, auf welche technischen Teile und auf welche abstrakten Risiken die Prüfung einer KI-Anwendung fokussieren sollte. In Abschnitt 2.1 werden deshalb der formale Aufbau sowie der Lebenszyklus einer KI-Anwendung in abstrahierter Form erläutert, die zur Ausrichtung innerhalb des Katalogs herangezogen werden können und ein einheitliches Verständnis über den Prüfgegenstand schaffen sollen.

Auf inhaltlicher Ebene deckt der KI-Prüfkatalog sechs Dimensionen der Vertrauenswürdigkeit ab, die sich wiederum in verschiedene Risikogebiete unterteilen. Diese thematische Strukturierung, entlang derer insbesondere Qualitätskriterien und risikoabschwächende Maßnahmen gegliedert sind, wird in Abschnitt 2.2 erläutert. Außerdem wird ein inhaltlicher Überblick über die verschiedenen Dimensionen und Risikogebiete gegeben.

Im letzten Abschnitt dieses Kapitels wird dargestellt, wie sich der Katalog konkret als Grundlage zur Prüfung anwenden lässt. Dazu wird der risikobasierte Prüfansatz des Katalogs beschrieben. Insbesondere werden Logik und Abfolge der im Prüfkatalog entwickelten Systematik im Detail erläutert, angefangen bei der Identifikation von KI-Risiken über die Herleitung von Qualitätskriterien bis hin zur abschließenden Beurteilung der KI-Anwendung.

---

17 Eine mithilfe des Katalogs betrachtete KI-Anwendung wird im Folgenden als Prüfgegenstand bezeichnet. Mit dieser Begrifflichkeit soll jedoch nicht ausgeschlossen werden, dass der Katalog neben der Beurteilung von KI-Anwendungen auch bei der Entwicklung bzw. zur Verbesserung von KI-Anwendungen herangezogen werden kann.

18 Der Prüfansatz dieses Katalogs wird auch erläutert in: Poretschkin, M., Mock, M., Wrobel, S.. Zur Systematischen Bewertung der Vertrauenswürdigkeit von KI-Systemen. In: D. Zimmer (Hrsg.), Regulierung für Algorithmen und Künstliche Intelligenz (in Erscheinung).

2.1 Prüfgegenstand

Zur Orientierung innerhalb des KI-Prüfkatalogs beschreibt dieser Abschnitt in allgemeiner Form, und somit auf einem hohen Abstraktionsgrad, zunächst den formalen Aufbau einer KI-Anwendung, indem es diesen in verschiedene funktionale Komponenten zerlegt. Je nach Art der eigentlichen Anwendung können diese unterschiedlich komplex oder umfangreich ausfallen. Neben dieser funktionalen Sicht wird auch der Lebenszyklus einer solchen Anwendung in Abschnitt 2.1.2 beleuchtet. Beide Blickwinkel werden innerhalb des Katalogs holistisch durch verschiedene Maßnahmenkategorien, vgl. Abschnitt 2.3.4, abgebildet. Hierbei betrifft eine Trennung nach Einbettung und KI-Komponente den Aufbau, während Daten und Betrieb grob zwischen verschiedenen Stadien des Lebenszyklus unterscheiden. Wie bei allen komplexen Systemen sind die Trennungen nicht notwendigerweise scharf, so können Daten beispielsweise auch während des Betriebs aufgenommen werden.

2.1.1 Struktur einer KI-Anwendung

Ein erster Schritt bei der Diskussion und Beurteilung einer KI-Anwendung besteht darin, ihren formalen Aufbau zu spezifizieren sowie den Prüfgegenstand abzugrenzen. Während der **KI-Steckbrief (ST)** einen ersten Überblick über die KI-Anwendung ermöglicht, werden im Rahmen der Prüfung, insbesondere in der Dimension Verlässlichkeit, detailliertere Dokumentationen des Prüfgegenstands angefordert. Ziel des vorliegenden Abschnitts ist es, ein einheitliches Verständnis von Begriffen bezüglich der Struktur einer KI-Anwendung zu schaffen.

Der KI-Prüfkatalog ist auf KI-Anwendungen ausgerichtet, die durch Maschinelles Lernen (ML) realisiert sind. Zwar spielen heutzutage im Bereich ML insbesondere Neuronale Netze eine wichtige Rolle, doch ist der Prüfkatalog nicht auf diese Technologie beschränkt. Auch andere Methoden, wie etwa Entscheidungsbäume oder Stützvektorverfahren, die ggf. sogar besser als Neuronale Netze Anforderungen in Hinblick auf Transparenz oder Sicherheit erfüllen, können mithilfe des Prüfkatalogs untersucht werden.

Das ML-Modell bildet den Kern der KI-Komponente, welche zusätzlich um Vor- und Nachbereitungsschritte der Ein- und Ausgaben des ML-Modells ergänzt ist. Das ML-Modell sowie die KI-Komponente sind mathematische Objekte, die als Grundlage einer Funktionalität im Sinne eines Eingabe-Ausgabe-Mappings dienen können, wie beispielsweise einer Objekterkennung oder der Bewertung von Bewerber*innen. In praktischen Anwendungsfällen wird das Eingabe-Ausgabe-Mapping meist im Zusammenspiel mit weiteren Komponenten der Einbettung durchgeführt, die etwa zur Protokollierung oder Überwachung der Eingaben, beispielsweise durch ein regelbasiertes (nicht-ML) Expertensystem, dienen können. Das derart in einem gegebenen Einsatzkontext ausübbare Eingabe-Ausgabe-Mapping wird in diesem Katalog als KI-Anwendung bezeichnet. Der Begriff der KI-Anwendung sowie ihre funktionalen Zusammenhänge, die die verschiedenen Verarbeitungsschritte der Eingabedaten repräsentieren, sind in abstrahierter Form in Abbildung 1 dargestellt und werden im danach folgenden Abschnitt näher erläutert.

Ist eine KI-Anwendung Teil eines möglicherweise nicht nur auf KI-Technologien basierenden Gesamtsystems, so sollten deren Grenzen innerhalb des Gesamtsystems klar definiert sein. Zum Beispiel kann eine ML-basierte Objekterkennung (KI-Anwendung) in ein autonomes Fahrzeug, in eine Drohne oder in ein Geländeüberwachungssystem (als Gesamtsystem) integriert sein. Die Abgrenzung einer KI-Anwendung innerhalb eines Gesamtsystems richtet sich maßgeblich nach ihrer Funktionalität. Im Beispiel einer KI-basierten Fußgängererkennung (KI-Anwendung) innerhalb eines autonomen Autos (Gesamtsystem) würden Softwaremodule, die etwa Konsistenzchecks auf den Ausgaben der KI-Komponente durchführen, zur KI-Anwendung zählen; andere Komponenten, die beispielsweise die Gesamtplanung einer Fahrtroute vornehmen, jedoch nicht.

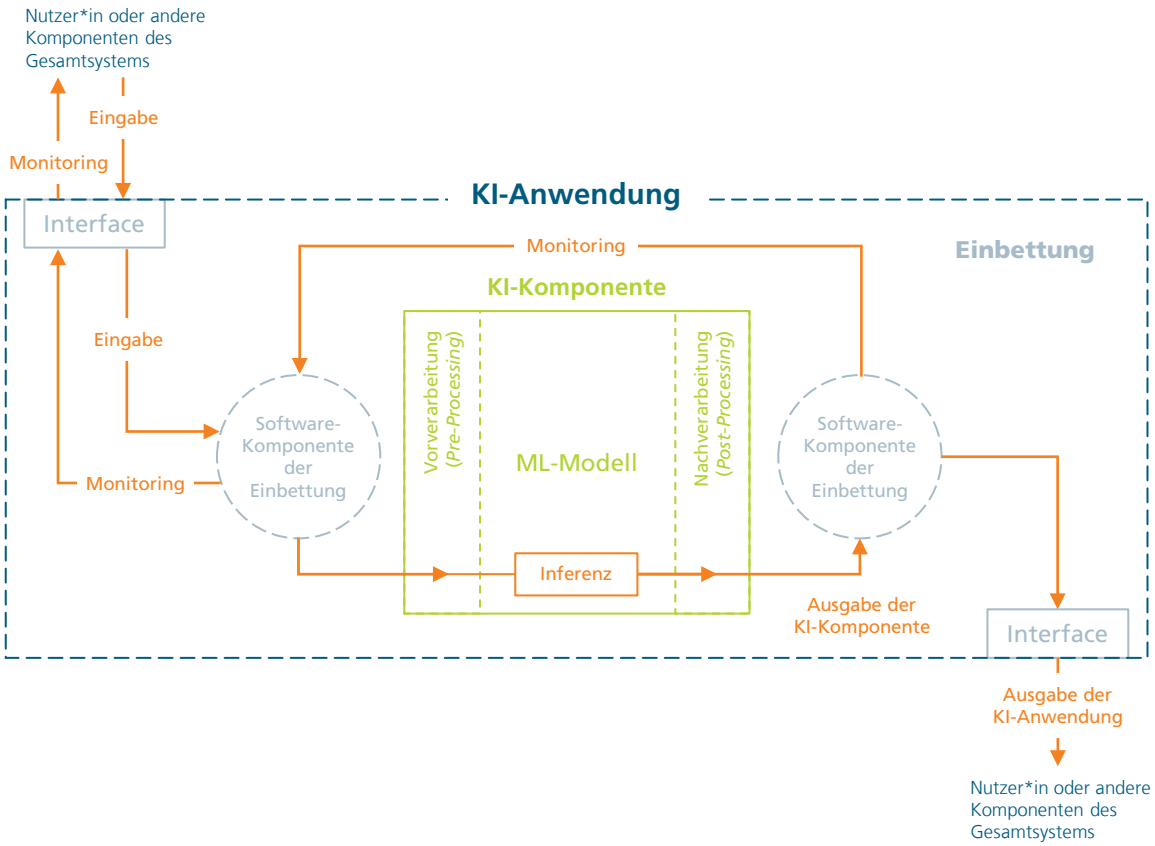

Abbildung 1: Formaler Aufbau einer KI-Anwendung

(ML-)Modell: Das ML-Modell ist ein mathematisches, abstraktes Objekt, das durch ein maschinelles Lernverfahren erstellt wurde und dazu dient, eine Problemstellung im Sinne einer Eingabe-Ausgabe-Relation zu lösen. Im Beispiel eines Neuronalen Netzes besteht das Modell u. a. aus einer Auflistung an Hyperparametern, gelernten Parametern sowie einer Beschreibung ihrer Interaktion zur Laufzeit (Architektur). Das ML-Modell liefert die funktionale Grundlage der KI-Anwendung. Beispielsweise kann ein Modell als Grundlage für die Ausübung einer Klassifikationsaufgabe dienen, wobei die Eingabe den zu klassifizierenden Gegenstand repräsentiert und die Ausgabe des Modells eine Aussage über seine Klasse trifft. In einigen Fällen, etwa für generative Modelle, kann die Eingabe (im Beispiel meist Zufallszahlen) für die eigentliche Funktion von untergeordneter Bedeutung sein.

KI-Komponente: Die KI-Komponente setzt sich zusammen aus dem ML-Modell sowie gegebenen modellspezifischen Verfahren zur Datenvorverarbeitung (*Pre-Processing*) und zur Nachverarbeitung der Modell-Ausgaben (*Post-Processing*). Somit ist auch die KI-Komponente ein mathematisches Objekt.

Anmerkung: Zur Vereinfachung wird im Prüfkatalog die Annahme getroffen, dass, sofern nicht explizit anders angegeben[19], eine KI-Anwendung genau eine KI-Komponente enthält.

19 Beispielsweise können Ensemble-Methoden als ein ML-Modell angesehen werden und fallen somit ausdrücklich in den Geltungsbereich des Prüfkatalogs.

Einbettung: Die KI-Komponente ist in der Regel mit weiteren (klassischen) Software-Modulen und technischen Komponenten verknüpft, etwa zur Speicherung von Daten oder zur Umsetzung physischer Reaktionen auf Ausgaben der KI-Komponente. Als Einbettung wird die Gesamtheit derjenigen umgebenden Komponenten bezeichnet, die sich unmittelbar auf die Funktions- und Wirkungsweise der KI-Komponente beziehen. Dazu gehören beispielsweise Software-Module, die die KI-Komponente aktivieren und ihre Ergebnisse weiterverarbeiten. Insbesondere zählen klassische Komponenten, die die Funktionalität der KI-Komponente nach außen sichtbar machen und eine Interaktion ermöglichen (siehe **Interface**) zur Einbettung. Darüber hinaus können Software-Module der Einbettung auch dazu dienen, ein Versagen der KI-Komponente festzustellen und abzufangen (in Abbildung 1 als *Monitoring* bezeichnet).

Interface/Schnittstelle: Als Interface, zu Deutsch Schnittstelle, wird der Teil der Einbettung bezeichnet, der eine Interaktion der KI-Anwendung mit der Außenwelt, wie etwa Nutzer*innen oder anderen Komponenten innerhalb eines Gesamtsystems, ermöglicht. Je nach Funktion und Gestaltung einer KI-Anwendung bietet das Interface verschiedene Interaktionsmöglichkeiten. In der Regel beinhalten diese zum einen, dass Inputdaten (z. B. Anfragen von Nutzer*innen) aufgenommen werden; zum anderen, dass die Ergebnisse/Ausgaben der KI-Komponente nach außen kommuniziert bzw. abrufbar gemacht werden.

KI-Anwendung: Als KI-Anwendung wird das Eingabe-Ausgabe-Mapping in einem gegebenen Einsatzkontext bezeichnet, das auf der implementierten KI-Komponente basiert. Der KI-Prüfkatalog nimmt insbesondere keine isolierte Betrachtung des ML-Modells oder der KI-Komponente vor. Hingegen untersucht er, ob die einzelnen Verarbeitungsschritte bis hin zu den Ergebnissen, die durch die KI-Komponente im Zusammenspiel mit weiteren Komponenten der Einbettung erzeugt werden, für den gegebenen Einsatzkontext sinnvoll und angemessen sind, d. h., ob sie beispielsweise in ausreichendem Maße frei von Fehlern und Diskriminierung oder sicher vor Angriffen und Manipulation sind. Prüfgegenstand des Katalogs sind somit nicht die mathematischen Konzepte an sich, die der KI-Komponente zugrunde liegen, sondern die Funktionalität im Sinne des Eingabe-Ausgabe-Mappings, das basierend auf der KI-Komponente (als funktionaler Grundlage) in einem gegebenen Einsatzkontext ausgeübt wird. Hierbei kann eine KI-Anwendung ein alleinstehendes System sein, das beispielsweise direkt mit Nutzer*innen kommuniziert oder sie kann in ein (IT-)Gesamtsystem oder Produkt integriert sein.

Anmerkung 1: Im Prinzip ist die KI-Anwendung ein abstrakter Prüfgegenstand. Jedoch kann sie in der Regel (und insbesondere im Fall komplexer Neuronaler Netze) nicht losgelöst von ihrer teilweise auch physischen Implementierung betrachtet werden. Insbesondere setzen die im Prüfkatalog beschriebenen Testmaßnahmen voraus, dass die KI-Anwendung implementiert oder sogar in der realen Einsatzumgebung bzw. im Endprodukt installiert und funktionsfähig ist. Der Fokus des KI-Prüfkatalogs liegt jedoch nicht auf den Herausforderungen bei der Implementierung der KI-Anwendung, sondern darauf, ob die KI-Komponente als Grundlage zur Ausübung der gewünschten Funktionalität geeignet und vertrauenswürdig ist.

Anmerkung 2: Sofern nicht explizit anders angegeben, wird in diesem Prüfkatalog der Fall, dass eine KI-Anwendung auf mehreren KI-Komponenten, d. h. insbesondere auf mehreren strukturell verschiedenen ML-Modellen, basiert, nicht betrachtet. Zwar können verschiedene KI-Komponenten mithilfe des KI-Prüfkatalogs separat betrachtet werden, jedoch wird deren Zusammenspiel im Prüfkatalog nicht gesondert behandelt.

2.1.2 Lebenszyklus einer KI-Anwendung

Neben ihrer Struktur ist der Lebenszyklus einer KI-Anwendung ein zentraler Ansatzpunkt zur Identifikation und Mitigation von KI-Risiken und sollte daher vollumfänglich bei einer Qualitätsbewertung betrachtet werden. Eine zentrale Bedeutung im Lebenszyklus einer KI-Anwendung, die durch Maschinelles Lernen realisiert ist, kommt den Daten zu, da die Funktionalität der Anwendung meist direkt aus Daten abgeleitet wird. Insbesondere ist der Lebenszyklus einer KI-Anwendung enger verbunden mit Prozessen des Data Mining als mit Entwicklung und Betrieb klassischer (vollständig von Menschen entworfener und programmierter)

IT-Systeme. Der CRISP-DM-Standard[20], der Data Mining als Prozess – angefangen bei der Zielsetzung und Datenauswahl über die Anwendungsentwicklung bis hin zum Betrieb – betrachtet, kann im Wesentlichen auch auf die häufig komplexeren Modellklassen der hier betrachteten KI-Anwendungen, etwa tiefe Neuronale Netze, übertragen werden. Dies ist in Abbildung 2 vereinfacht dargestellt.

Abbildung 2: Abstrahierter Lebenszyklus einer KI-Anwendung

Insbesondere der Auswahl und Aufbereitung von Daten kommt eine entscheidende Rolle zu, da Eigenschaften einer KI-Anwendung (z. B. die gelernten Gewichte) und damit auch ihre Funktionalität aus der Optimierung ausgehend von Daten resultieren. Verglichen mit ursprünglichem Data Mining lösen komplexe maschinelle Lernverfahren in vielen Fällen das manuelle Vorverarbeiten von Daten, etwa das Herausarbeiten relevanter Dateneigenschaften (Features), weitestgehend ab. Aus diesem Grund müssen Risiken hinsichtlich der Daten stärker in den Blick genommen und angemessene Maßnahmen zur Erfüllung von Anforderungen an die Daten in der Qualitätsbewertung eingebracht werden. Diese können sich beispielsweise auf die Fairness, insbesondere die Möglichkeit eines inhärenten Bias, oder die Qualität und Eignung von Daten, etwa mit Blick auf eine hinreichende Abdeckung oder die Qualität von Labels, beziehen.

Die Entwicklung, welche die zweite Phase im Lebenszyklus manifestiert, umfasst Aspekte des Designs, der Modellbildung sowie des Testens. Diesbezüglich orientiert sich der Prüfkatalog an der in Abschnitt 2.1.1 beschriebenen Aufteilung zwischen KI-Komponente und ihrer umgebenden Einbettung. Durch deren Design, beispielsweise durch die Auswahl einer bestimmten Modellarchitektur oder durch die Entscheidung für Einbettungskomponenten, die etwa Konsistenzchecks auf Ausgaben der KI-Komponenten durchführen oder zu einem gewissen Grad Redundanz bei Ausfall der KI-Komponente herstellen, können angestrebte Qualitätseigenschaften einer KI-Anwendung begünstigt oder gar erreicht werden. Essenziell für die Qualität von KI-Anwendungen ist ferner das Training des ML-Modells, welches in Abbildung 3 anhand eines einfachen Standardablaufs erläutert wird. Als Training wird der Vorgang bezeichnet, bei dem ein Modell mithilfe eines Maschinellen Lernverfahrens erstellt wird. Hierzu werden die Parameter des Modells, auch Gewichte genannt, durch eine von sogenannten Hyperparametern bestimmte Optimierung ausgehend von Trainingsdaten ermittelt. Hierbei werden nach meist zufälliger Initialisierung der Gewichte Trainingsdaten in das Modell eingegeben. Anschließend werden die damit berechneten Ergebnisse, im Fall überwachter ML-Verfahren unter Verwendung der zu den Daten zugehörigen Labels (*Ground Truth*), anhand einer Verlustfunktion (auch Loss-Funktion genannt) bewertet. Hierbei misst die Loss-Funktion die Diskrepanz zwischen den Ergebnissen des Modells und der *Ground Truth*. Die Bewertung der Loss-Funktion wird durch Anpassung der Gewichte iterativ optimiert. Da die Hyperparameter hierbei einen starken Einfluss auf die final ermittelten Parameter haben können, sollte zusätzlich mithilfe eines weiteren (Validierungs-)Datensatzes verglichen werden, welche Hyperparameterkonfiguration am besten für die Lösung der gegebenen Problemstellung geeignet ist. Neben der Verlustfunktion, die Objekt der Optimierung ist, sollten zur Beurteilung des ML-Modells weitere Performanz-Metriken herangezogen werden. Insgesamt hängt die Güte der KI-Komponente von einer Vielzahl an Faktoren ab, darunter die Wahl der Datenbasis, des Lernalgorithmus inklusive der Verlustfunktion, sowie der Hyperparameter. Entsprechend wichtig für die Sicherstellung ihrer Qualität sind Tests, wobei auf die Wahl geeigneter Testdaten geachtet werden sollte. Diese Daten dürfen nicht den

20 Für eine ausführliche Beschreibung des Cross-Industry Standard Process for Data Mining (CRISP-DM) siehe: Shearer, C. (2000). The CRISP-DM model: The new blueprint for data mining. Journal of Data Warehousing, 5(4):13–22.

Trainings- oder Validierungsdaten entsprechen, da das ML Modell direkt für diese optimiert wurde. Tests können sowohl das Verhalten der KI-Anwendung angesichts der im Betrieb zu erwartenden Eingabedaten bewerten, als auch dazu dienen, gezielt nach Schwachstellen des Modells zu suchen.

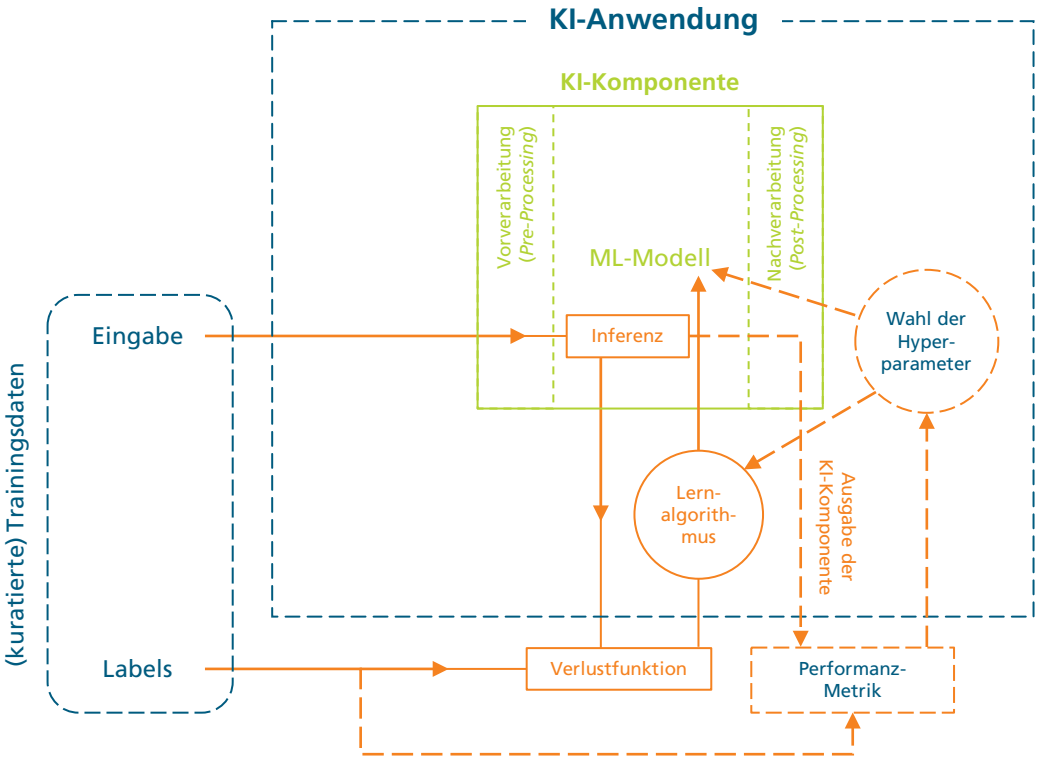

Abbildung 3: Training des ML-Modells einer KI-Anwendung.

Im laufenden Betrieb ergeben sich anschließend zwei unterschiedliche Blickwinkel, die im Rahmen einer Prüfung berücksichtigt werden sollten. Zum einen sollten die Eingriffs- und Feedbackmöglichkeiten menschlicher Nutzer*innen und Betroffenen sichergestellt sein. Zum anderen hat der datengetriebene Ansatz von KI-Anwendungen Risiken des *Concept* und *Model Drift* zur Folge, gegen welche entsprechende Vorkehrungen getroffen werden müssen. Bei *Concept Drift* handelt es sich um das Risiko, dass sich die Beschaffenheit der Eingabedaten oder Rahmenbedingungen ändern und die KI-Anwendung daher im Laufe des Betriebs nicht mehr den Anforderungen entspricht. *Model Drift* hingegen beschreibt den Umstand, dass die KI-Anwendung (genauer das ML-Modell) im laufenden Betrieb weiterlernt und aufgrund seiner resultierenden Veränderung nicht mehr den Anforderungen genügt.

2.2 Dimensionen der Vertrauenswürdigkeit

Während ML-Technologien vielfältige Möglichkeiten eröffnen, bergen sie andererseits aufgrund ihrer Komplexität, Dynamik und Intransparenz neue Risiken. Insbesondere ist ein Großteil der KI-spezifischen Risiken nicht durch die bestehenden Prüf- und Zertifizierungsverfahren für klassische IT-Systeme abgedeckt. Der KI-Prüfkatalog liefert einen Ansatz zur strukturierten Evaluation von KI-Anwendungen, wobei er jedoch nicht den Anspruch hat, bestehende Prüfverfahren oder Standards für (klassische) IT-Systeme beispielsweise zur IT-Sicherheit für KI-Anwendungen neu zu schreiben bzw. abzulösen. Vielmehr soll der KI-Prüfkatalog als kompatible Ergänzung zu bestehenden Prüf- und Zertifizierungsverfahren dienen, mit dem Ziel, die dargestellte Lücke zu schließen. Insbesondere liegt der Fokus des KI-Prüfkatalogs auf den KI-spezifischen Risiken von KI-Anwendungen.

Die Frage, welche KI-spezifischen Risiken durch eine KI-Prüfung abgedeckt werden müssen und nach welchen Kriterien diese zu bewerten sind, ist seit geraumer Zeit Gegenstand intensiver gesellschaftlicher und wissenschaftlicher Debatten. Aus den verschiedenen Beiträgen zu dieser Diskussion – darunter besonders prominent die Kernanforderungen der HLEG[21] – haben sich sechs Themenfelder herauskristallisiert, die im Folgenden als Dimensionen der Vertrauenswürdigkeit bezeichnet werden: Fairness, Autonomie und Kontrolle, Transparenz, Verlässlichkeit, Sicherheit sowie Datenschutz.

Diese sechs Dimensionen der Vertrauenswürdigkeit bilden die grundlegende thematische Gliederung des KI-Prüfkatalogs und dienen insbesondere als strukturierter und granularer Ansatz zur Herleitung von Qualitätskriterien. Die Darstellung der Dimensionen in diesem Katalog orientiert sich an dem Whitepaper »Vertrauenswürdiger Einsatz von Künstlicher Intelligenz«[22], das in einem interdisziplinären Dialog von Informatiker*innen, Jurist*innen und Philosoph*innen erarbeitet wurde. Für eine umfassende Diskussion der sechs Dimensionen wird auf das Whitepaper verwiesen.

Der inhaltliche Fokus der sechs Dimensionen ist jeweils auf solche Risiken ausgerichtet, die von dem in der KI-Anwendung implementierten maschinellen Lernverfahren ausgehen oder zumindest im unmittelbaren Zusammenhang mit dessen Funktionalität stehen. Neben Risiken, die erst durch den Einsatz maschineller Lernverfahren aufkommen, sind dabei insbesondere auch solche Risiken inbegriffen, die zwar bereits in bestehenden Standards adressiert werden, aber durch den Einsatz von KI signifikant an Bedeutung gewinnen, wie etwa der Schutz von Daten. In diesem Fall erläutert der Prüfkatalog die neuartigen Risikofaktoren und bringt KI-spezifische Maßnahmen an, mit denen diese zusätzlich zu den bekannten klassischen Maßnahmen abgeschwächt werden können. Diesem Fokus entsprechend fallen andererseits Analysen der Code-Qualität oder etwa der Hardware-Sicherheit explizit aus der Betrachtung des Prüfkatalogs heraus. Auch werden in der Dimension: Sicherheit solche Sicherheitsrisiken, die von der Einbettung ausgehen, oder die gar bei fehlerfreiem Betrieb der KI-Anwendung bestehen, nicht betrachtet, da sie KI-unspezifisch sind.

Die Dimensionen der Vertrauenswürdigkeit werden im Prüfkatalog weiter in Risikogebiete unterteilt. Diese haben den Zweck, verwandte Risiken innerhalb einer Dimension zu bündeln, die durch ähnliche Maßnahmen abgeschwächt werden können. Dementsprechend unterscheiden sich die Risikogebiete etwa in Bezug auf betrachtete Fehlerursachen oder Angriffsszenarien. So werden in der Dimension Verlässlichkeit Fragestellungen in Hinblick auf die Performanz der KI-Anwendung unter normalen Betriebsbedingungen, auf den Umgang der KI-Anwendung mit potenziell auftretenden Störungen sowie auf das Weiterlernen des ML-Modells in separaten Risikogebieten behandelt. Insbesondere werden die Qualitätskriterien zur Beurteilung der Vertrauenswürdigkeit auf Ebene der Risikogebiete hergeleitet.

---

21 Die High-Level Expert Group on AI (HLEG) ist eine von der Europäischen Kommission einberufene Expertenkommission für Künstliche Intelligenz. Sie hat sieben Kernanforderungen für vertrauenswürdige KI formuliert, siehe dazu: High-Level Expert Group on AI (HLEG). (April 2019). Ethics Guidelines on trustworthy AI. Europäische Kommission. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (letzter Aufruf: 21.06.2021).

22 Poretschkin, M., et al. (2019). Vertrauenswürdiger Einsatz von Künstlicher Intelligenz. Sankt Augustin: Fraunhofer-Institut für Intelligente Analyse und Informationssysteme IAIS. https://www.ki.nrw/wp-content/uploads/2020/03/Whitepaper_KI-Zertifizierung.pdf (letzter Aufruf: 18.06.2021) Anmerkung: In dem Whitepaper werden die Dimensionen als »Handlungsfelder« bezeichnet.

Im Folgenden werden die sechs Dimensionen der Vertrauenswürdigkeit und die darin enthaltenen Risikogebiete genauer vorgestellt.

Anmerkung: Zur besseren Übersichtlichkeit wird jeder Dimension und jedem Risikogebiet ein Kürzel zugeordnet. Dieses wird später dazu verwendet, um u. a. Kriterien und Maßnahmen mit einer eindeutigen Kennung, einem alphanumerischen Identifier, zu versehen (siehe auch Abschnitt 2.3).

Beispiel: Die Dimension Fairness hat das Kürzel »FN«, das Risikogebiet »Beherrschung der Dynamik« in dieser Dimension das Kürzel »BD«. Das erste Kriterium in diesem Risikogebiet hat im Prüfkatalog die Kennung [FN-R-BD-KR-01], wobei »R« für Risikogebiet und »KR« für Kriterium steht.

2.2.1 Dimension Fairness

Fairness (FN)	Die Dimension Fairness soll sicherstellen, dass die KI-Anwendung nicht zu ungerechtfertigter Diskriminierung führt. Typische Ursachen hierfür stellen unausgewogene (mit Bias behaftete) Trainingsdaten oder auch die statistische Unterrepräsentation von Personengruppen dar, welche zu einer verringerten Qualität der KI-Anwendung in Bezug auf diese Gruppen führen können.		
	Risikogebiete	Fairness (FN)	Dieses Risikogebiet adressiert das Risiko, dass die KI-Anwendung während der Entwicklung unfaires bzw. diskriminierendes Verhalten gegenüber Nutzer*innen oder Betroffenen lernt.
		Beherrschung der Dynamik (BD)	Dieses Risikogebiet behandelt Risiken hinsichtlich Fairness, die sich durch Veränderungen der Rahmenbedingungen oder Änderungen im Nutzerverhalten ergeben.

2.2.2 Dimension Autonomie und Kontrolle

Autonomie und Kontrolle (AK)	Diese Dimension hebt auf zwei Dinge ab: zum einen die Autonomie der KI-Anwendung und zum anderen die Autonomie des Menschen. Einerseits ist hier zu beurteilen, welcher Grad an Autonomie für die Anwendung angemessen ist (z. B. *Human-in/on/out-of-the-Loop*[23]). Andererseits wird untersucht, ob der Mensch durch die KI-Anwendung angemessen unterstützt wird und ausreichend Handlungsspielraum in der Interaktion mit der KI-Anwendung erhält.		
	Risikogebiete	Angemessene und verantwortungsvolle Gestaltung der Aufgabenverteilung zwischen Mensch und KI-Anwendung (GE)	Dieses Risikogebiet behandelt Risiken, die sich aus Einschränkungen der Nutzerautonomie bzw. unangemessener Autonomie der KI-Anwendung ergeben.
		Sicherstellung der Informiertheit und Befähigung von Nutzer*innen und Betroffenen (IB)	Dieses Risikogebiet adressiert Risiken, die dadurch entstehen, dass Nutzer*innen und Betroffene unzureichend über die KI-Anwendung, deren Nutzung sowie die damit verbundenen Risiken aufgeklärt werden.

2.2.3 Dimension Transparenz

Transparenz (TR)	Unter diesem Oberbegriff sind Aspekte der Nachvollziehbarkeit, Reproduzierbarkeit und Erklärbarkeit subsumiert. Die Dimension Transparenz untersucht insbesondere, ob die grundlegende Funktionsweise der KI-Anwendung für Nutzer*innen und Experten*innen angemessen nachvollziehbar ist, und ob Ergebnisse der KI-Anwendung reproduziert und ggf. begründet werden können.		
	Risikogebiete	Transparenz gegenüber Nutzer*innen und Betroffenen (NB)	Dieses Risikogebiet befasst sich mit Risiken, die dadurch entstehen, dass Entscheidungen und Auswirkungen der KI-Anwendung gegenüber Nutzer*innen und Betroffenen nicht hinreichend erklärt werden können.
		Transparenz für Expert*innen (EX)	Dieses Risikogebiet widmet sich Risiken, die dadurch entstehen, dass das Verhalten der KI-Anwendung von einem*einer Expert*in nicht hinreichend verstanden und nachvollzogen werden kann.
		Auditfähigkeit (AF)	Dieses Risikogebiet behandelt Risiken, die dadurch entstehen, dass die Entwicklung sowie die im Einzelfall ausgeführten Vorgänge im Betrieb der KI-Anwendung nicht hinreichend dokumentiert und belegt sind.
		Beherrschung der Dynamik (BD)	Dieses Risikogebiet behandelt Risiken, die dadurch entstehen, dass sich Anforderungen an die Transparenz oder die implementierten Transparenzverfahren selbst ändern.

23 Für eine Erläuterung der Autonomiegrade siehe: Nothwang, W., et al. (2016). The Human Should be Part of the Control Loop? In 2016 Resilience Week (RWS), pp. 214-220, IEEE https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7573336 (letzter Zugriff: 22.06.2021) Anmerkung: Der Autonomiegrad, der im Paper »Complete Autonomy« genannt wird, ist in diesem Katalog als »Human-out-of-the-Loop« bezeichnet.

2.2.4 Dimension Verlässlichkeit

Verlässlichkeit (VE)	Diese Dimension bezieht sich vornehmlich auf die Qualität der KI-Komponente und beurteilt u. a. deren Robustheit, d. h. die Konsistenz ihrer Ausgaben unter kleinen Veränderungen der Eingabedaten. Neben der Performanz und Robustheit der KI-Komponente wird auch deren potenzielle Ausgabe(un)sicherheit betrachtet.		
	Risikogebiete	Verlässlichkeit im Regelfall (RE)	Dieses Risikogebiet behandelt das Risiko fehlerhafter Vorhersagen durch die KI-Komponente auf regulären Eingabedaten.
		Robustheit (RO)	Dieses Risikogebiet adressiert Risiken, die sich bei störungsbehafteten oder manipulierten Eingabedaten ergeben, für die jedoch eine korrekte Verarbeitung durch die KI-Komponente beabsichtigt ist. Dabei werden sowohl qualitative als auch quantitative Abweichungen der Eingabedaten betrachtet, wie etwa Rauschen oder adversariale Beispiele.
		Abfangen von Fehlern auf Modellebene (AF)	Dieses Risikogebiet behandelt Risiken aus Eingabedaten, die nicht im Anwendungsbereich liegen und für die eine korrekte Bearbeitung durch die KI-Komponente nicht zu erwarten ist. Diese sollen durch eine Detektionsstrategie abgefangen werden.
		Einschätzung von Unsicherheit (UN)	Dieses Risikogebiet betrachtet Risiken, die sich daraus ergeben, dass keine Unsicherheitsschätzung bezüglich Ausgaben stattfindet, oder, dass diese unrealistisch bzw. unbrauchbar ist.
		Beherrschung der Dynamik (BD)	Dieses Risikogebiet befasst sich mit dem Risiko, dass das in der KI-Komponente implementierte ML-Modell aufgrund von unbeabsichtigten *Model Drifts* oder Veränderungen des Anwendungskontexts (*Concept Drift*) Einbußen an Performanz oder hinsichtlich anderer Anforderungen erfährt.

2.2.5 Dimension Sicherheit

Sicherheit (SI)	Diese Dimension adressiert sowohl Eigenschaften der Funktionalen Sicherheit als auch die Absicherung gegenüber Angriffen und Manipulationen der KI-Anwendung. Die Maßnahmen in dieser Dimension beziehen sich primär auf die Einbettung der KI-Komponente und umfassen unter anderem klassische Methoden der IT-Sicherheit.		
	Risikogebiete	Funktionale Sicherheit (FS)	Dieses Risikogebiet behandelt das Risiko unbeabsichtigter Personen- oder Sachschäden, die durch Fehlfunktion bzw. Ausfall der KI-Anwendung infolge mangelhaften Designs der Einbettung begünstigt oder gar verursacht werden.
		Integrität und Verfügbarkeit (IV)	Dieses Risikogebiet adressiert Risiken, die dadurch entstehen, dass die für den Betrieb der KI-Anwendung relevanten Daten verfälscht werden und dadurch die KI-Anwendung manipuliert wird und ggf. nicht mehr verfügbar ist.
		Beherrschung der Dynamik (BD)	Dieses Risikogebiet behandelt Risiken, die dadurch entstehen, dass neue Gefährdungen der genannten Risikogebiete auftreten oder etablierte Verfahren an Effektivität verlieren.

2.2.6 Dimension Datenschutz

Datenschutz (DS)	Diese Dimension bezieht sich auf den Schutz sensibler Daten im Kontext von Entwicklung und Betrieb einer KI-Anwendung. Dabei wird sowohl der Schutz personenbezogener Daten als auch von Geschäftsgeheimnissen adressiert.		
	Risikogebiete	Schutz personenbezogener Daten (PD)	Dieses Risikogebiet behandelt Risiken, die mit der nicht DSGVO-konformen Nutzung personenbezogener Daten durch die KI-Anwendung verbunden sind, sowie das Risiko der Re-Identifikation von Personen in einem Datensatz.
		Schutz geschäftsrelevanter Information (GI)	Dieses Risikogebiet adressiert Risiken, die dadurch entstehen, dass geschäftsrelevante Informationen durch die KI-Anwendung unerwünscht preisgegeben werden.
		Beherrschung der Dynamik (BD)	Dieses Risikogebiet behandelt die Risiken, dass neue Hintergrundinformationen etwa zur Erstellung eines Personenbezugs entstehen, oder dass sich die Anforderungen an die Verarbeitung von Daten durch die KI-Anwendung ändern.

2.3 Logik des Prüfverfahrens

Der Prüfkatalog bietet einen Leitfaden zur strukturierten Bewertung von KI-Anwendungen. In diesem Kapitel wird beschrieben, wie der Katalog konkret zur Durchführung einer KI-Prüfung verwendet werden kann. Durch seine risikobasierte Systematik stellt er dabei eine kompatible Ergänzung zu bestehenden Prüfverfahren dar, die den drängenden Bedarf der Prüfung hinsichtlich KI-spezifischer Risiken adressiert. Im Folgenden wird die risikobasierte Systematik des Prüfkatalogs beschrieben und es werden insbesondere die Logik sowie die einzelnen Schritte des Prüfverfahrens im Detail erläutert.

Der KI-Prüfkatalog kann Prüfer*innen in verschiedener Hinsicht bei der Bewertung von KI-Anwendungen unterstützen. Er dient etwa als Leitfaden zur Anfertigung einer technischen Dokumentation, in welcher Entwickler*innen die Vertrauenswürdigkeit einer KI-Anwendung strukturiert darlegen. Dazu werden unter Anleitung des Prüfkatalogs zunächst die identifizierten Risiken im spezifischen Einsatzkontext dokumentiert, nachfolgend die Qualitätskriterien, die zur Beurteilung der KI-Anwendung herangezogen werden und schließlich die technischen Vorkehrungen, Maßnahmen und Testergebnisse, mit denen die Erfüllung dieser Kriterien begründet wird. Wie im Verordnungsentwurf der Europäischen Kommission für Hochrisikosysteme gefordert, kann eine solche technische Dokumentation als Grundlage für eine Konformitätsbewertung der KI-Anwendung dienen. Hierbei beurteilt ein*eine Prüfer*in die Plausibilität, Vollständigkeit und Angemessenheit der Dokumentation und bewertet auf deren Basis die Vertrauenswürdigkeit der KI-Anwendung.

Neben diesem Vorgehen sind auch andere Verwendungsmöglichkeiten des Prüfkatalogs denkbar. Beispielsweise kann ein*eine Prüfer*in den Prüfkatalog als Hilfestellung nutzen, um entlang der darin aufgeführten Risiken und Testmaßnahmen die Prüfung einer gegebenen KI-Anwendung zu planen. Insbesondere kann ein*eine Prüfer*in anwendungsspezifische Tests durchführen, anstatt auf dokumentierte Testergebnisse des*der Entwickler*in zurückzugreifen. Solch unabhängige und tiefergehende Untersuchungen heben die Prüfung auf ein höheres *Assurance Level*.

Der Prüfansatz des Katalogs gliedert sich in zwei Phasen. Die erste Phase dient der Operationalisierung von Qualitätskriterien. Hierzu werden KI-Risiken identifiziert und analysiert, und darauf basierend anwendungsspezifische Qualitätskriterien hergeleitet. In der zweiten Phase wird eine Absicherungsargumentation für die KI-Anwendung entwickelt. Unter Würdigung ergriffener, risikoabschwächender Maßnahmen sowie durchgeführter Tests wird die Erfüllung der Qualitätskriterien begründet. Insgesamt entspricht die Logik des Prüfverfahrens einem *Top-Down and Bottom-Up Approach* zur Argumentation für die Vertrauenswürdigkeit einer KI-Anwendung. Die erste Phase (*Top-Down*) ist in Abbildung 4 dargestellt, die zweite Phase (*Bottom-Up*) in Abbildung 5. Wird der Prüfkatalog als Leitfaden zur Anfertigung einer Dokumentation der KI-Anwendung genutzt, so sind die Parallelogramme in den Abbildungen als empfohlene Dokumentationsschritte aufzufassen.

Die erste Phase des Prüfverfahrens entspricht einem *Top-Down*-Ansatz zur Operationalisierung von Qualitätskriterien. Zunächst wird im Rahmen einer Schutzbedarfsanalyse die Relevanz der einzelnen Dimensionen der Vertrauenswürdigkeit separat untersucht. Sofern diese nicht »gering« ist, erfolgt eine detaillierte Risikoanalyse entlang der den Dimensionen untergeordneten Risikogebieten. Hieraus ergeben sich Zielvorgaben. Anschließend wird eine Reihe qualitativer oder wo möglich quantitativer Kriterien definiert, anhand derer das Erreichen der Zielvorgaben bewertet werden kann. Hierzu zeigt der Prüfkatalog typische Kriterien auf.

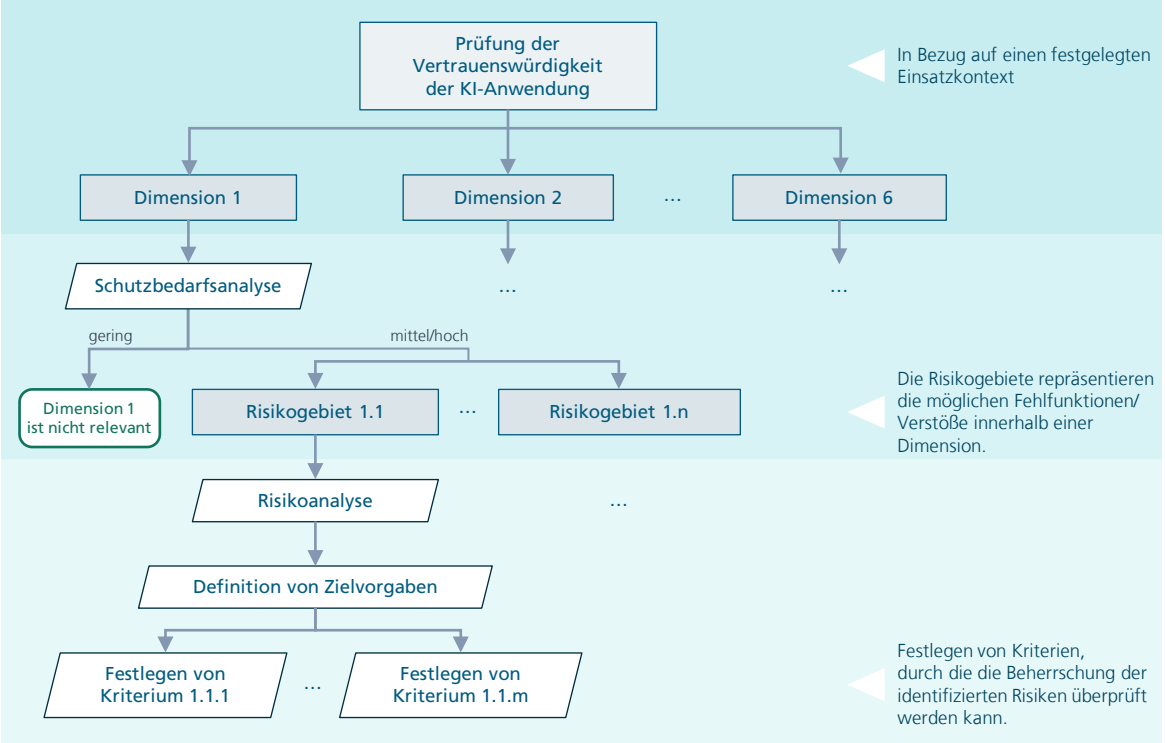

Abbildung 4: Top-Down – Risikobasierte Herleitung anwendungsspezifischer Qualitätskriterien. Hinweis: Die Parallelogramme stellen die Dokumentationsschritte dar.

Die zweite Phase folgt in entgegengesetzter Richtung einem *Bottom-Up*-Ansatz. Innerhalb eines jeden Risikogebiets werden Maßnahmen ergriffen und/oder dokumentiert, die auf eine Abschwächung von Risiken bzw. auf die Erfüllung der zuvor definierten Qualitätskriterien hinwirken. Diese Maßnahmen werden nach ihrem Ansatzpunkt bezüglich der KI-Anwendung unterschieden, d. h. ob sie die Daten, die KI-Komponente, ihre Einbettung oder später den laufenden Betrieb der KI-Anwendung betreffen. Der Prüfkatalog zeigt hierbei eine Vorgehensweise zur Abschwächung von Risiken auf. Basierend auf den ergriffenen Maßnahmen wird anschließend die Erfüllung der Kriterien argumentiert und es wird im Rahmen einer Gesamtbewertung begründet, bis zu welchem Grad relevante Risiken für das jeweilige Risikogebiet mitigiert sind. Eine vergleichbare Argumentation erfolgt auf Ebene der einzelnen Dimensionen sowie dimensionsübergreifend. Insbesondere bezüglich letzteren Schritts sind auch eventuell vorhandene Trade-Offs zwischen den Dimensionen zu berücksichtigen.

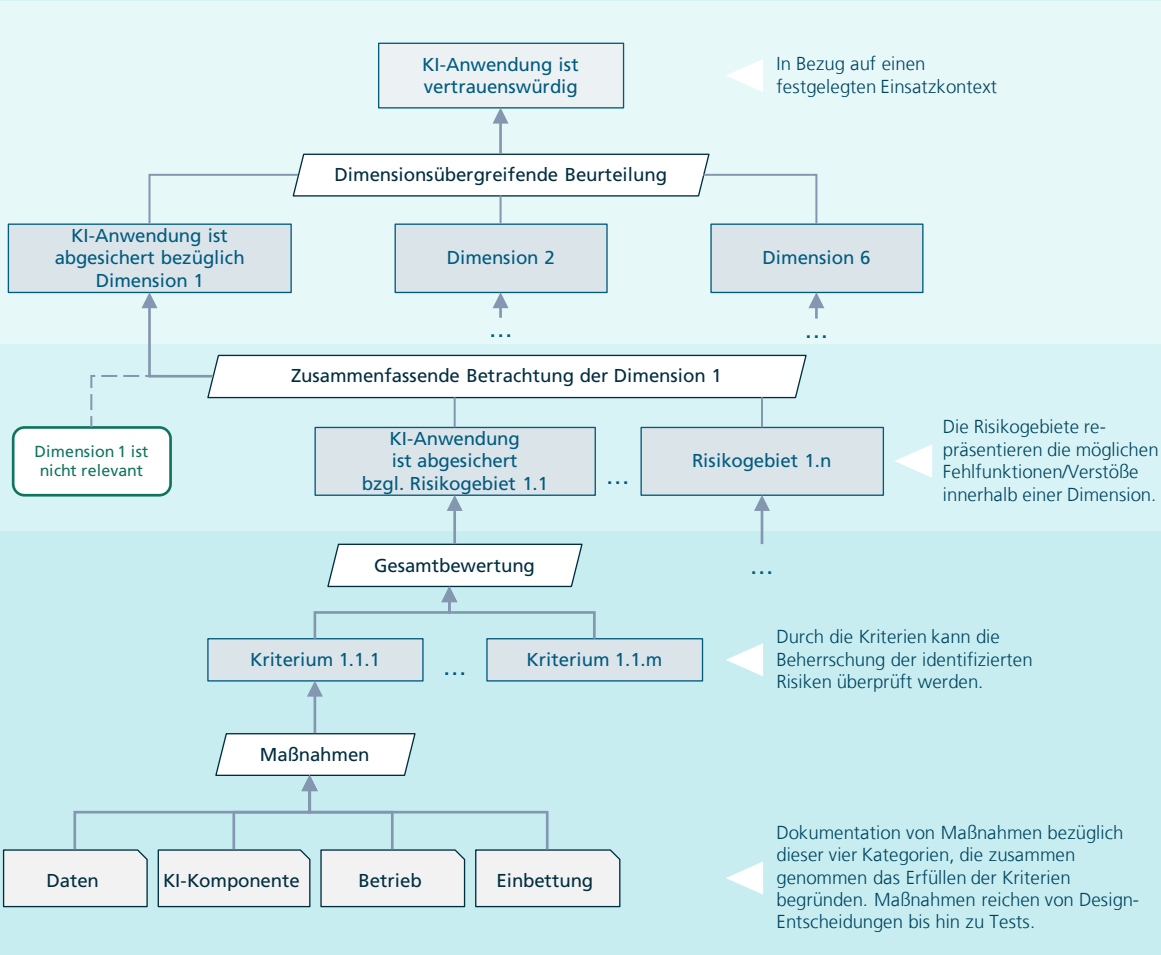

Abbildung 5: Bottom-Up – Argumentation für die Vertrauenswürdigkeit der KI-Anwendung basierend auf ergriffenen Maßnahmen. Hinweis: Die Parallelogramme stellen die Dokumentationsschritte dar.

Anmerkung: Wie in Abschnitt 2.2 bereits angemerkt, werden die einzelnen Schritte des Prüfverfahrens zur besseren Übersichtlichkeit mit Kennungen versehen. Eine Kennung in ihrer allgemeinen Form ist hierbei gegeben durch:

[Kapitel - Kategorie - ggf. nähere Bezeichnung der Kategorie - ggf. Aspekt - ggf. Nummerierung]

Beispiel 1: Zunächst wird in der Dimension Fairness **[FN]** eine Schutzbedarfsanalyse **[S]** durchgeführt. Diese hat die Kennung: **[FN-S]**

Beispiel 2: In der Dimension Fairness **[FN]** hat im Risikogebiet **[R]** Beherrschung der Dynamik **[BD]** das erste Kriterium **[KR]** die Nummer **[01]**: **[FN-R-BD-KR-01]**

Die Zusammensetzung der Kennungen wird zusätzlich durch die folgende Tabelle erklärt:

<table>
<tr><th>Kapitel</th><th>Kategorie</th><th>Aspekt</th></tr>
<tr><td>ST (KI-Steckbrief)</td><td colspan="2">B - Kürzel Themenbereich</td></tr>
<tr><td rowspan="6">Kürzel Dimension</td><td colspan="2">S (Schutzbedarfsanalyse)</td></tr>
<tr><td rowspan="4">R - Kürzel Risikogebiet</td><td>RI (Risikoanalyse und Zielvorgaben)</td></tr>
<tr><td>KR (Kriterien zur Zielerreichung)</td></tr>
<tr><td>MA (Maßnahmen)</td></tr>
<tr><td>BW (Gesamtbewertung)</td></tr>
<tr><td colspan="2">Z (Zusammenfassende Betrachtung)</td></tr>
<tr><td colspan="3">BV (Beurteilung der Vertrauenswürdigkeit)</td></tr>
</table>

Abbildung 6: Zusammensetzung und Bedeutung der Kennungen.

Nachfolgend werden die einzelnen Schritte und Aspekte des Prüfverfahrens im Detail vorgestellt.

2.3.1 Schutzbedarfsanalyse

Der KI-Prüfkatalog bietet ein strukturiertes Vorgehen zur Beurteilung einer KI-Anwendung in Hinblick auf sechs Dimensionen der Vertrauenswürdigkeit. Abhängig von Aufgabe und Einsatzkontext der KI-Anwendung ist es jedoch möglich, dass nicht alle sechs Dimensionen gleichermaßen relevant sind, um ihre Vertrauenswürdigkeit zu bewerten. Aus diesem Grund wird für jede Dimension zunächst eine sogenannte Schutzbedarfsanalyse durchgeführt. Ziel der Analyse ist es, in Anlehnung an die Vorgehensweise aus dem IT-Grundschutz, den sogenannten »Schutzbedarf« einer Dimension festzustellen. Dieser gleicht einer ersten Einschätzung der Relevanz dieser Dimension für die zu prüfende KI-Anwendung.

Die Ermittlung des Schutzbedarfs orientiert sich an den potenziellen Schäden[24], die durch eine Fehlfunktion bzw. eine Verletzung von Anforderungen in Hinblick auf die betrachtete Dimension entstehen können. Der Schutzbedarf wird in den Kategorien *gering*, *mittel* und *hoch* gemessen. Die einzige Ausnahme in der Ermittlung des Schutzbedarfs besteht in der Dimension Verlässlichkeit. Hier ist ein *geringer* Schutzbedarf ausgeschlossen. Dies rührt aus der Tatsache, dass ein geringer Schutzbedarf der Verlässlichkeit bedeuten würde, dass bei einer Fehlfunktion insbesondere finanzielle Schäden, etwa durch Verdienstausfall oder Rufschädigung, gar nicht oder nur in vernachlässigbarem Ausmaß drohen. Somit wäre die Qualität der KI-Anwendung vollkommen unkritisch und die Anwendung könnte, überspitzt formuliert, auch durch ein nicht-KI-basiertes System, das zufällige Ausgaben erzeugt, ersetzt werden. Hierbei ist außerdem zu berücksichtigen, dass für die Sinnhaftigkeit der anderen Dimensionen ein Mindestmaß an Funktionalität vorausgesetzt werden sollte. Aus diesem Grund scheint die Prüfung einer KI-Anwendung, die bezüglich Verlässlichkeit geringen Schutzbedarf hat, nicht erforderlich.

Wird für eine Dimension ein geringer Schutzbedarf festgestellt, so kann von einer näheren Betrachtung dieser Dimension abgesehen werden, da keine nennenswerten Risiken zu adressieren sind. Beispielsweise ist die Dimension Fairness zu vernachlässigen, wenn die KI-Anwendung weder personenbezogene Daten verarbeitet noch menschliche Nutzer*innen wesentlich durch die Ergebnisse betroffen sind. Dieser Fall kann etwa bei

24 Hierzu zählen zum einen (materielle) Sach- und Personenschäden sowie immaterielle Schäden an Nutzer*innen oder Betroffenen, die jeweils unmittelbar durch die KI-Anwendung verursacht werden. Zum anderen sind auch indirekte Schäden, die mit Fehlfunktion der KI-Anwendung einhergehen, wie etwa finanzielle Einbußen einer Organisation aufgrund von Rufschädigung, zu berücksichtigen.

KI-Anwendungen, die im industriellen Kontext eingesetzt werden, eintreten. Durch das Übergehen von Dimensionen mit geringem Schutzbedarf wird die Prüfung passgenau und damit effizient gestaltet. Umgekehrt kann ein hoher Schutzbedarf ein besonderes Augenmerk auf die entsprechende Dimension im Rahmen der Prüfung erforderlich machen.

2.3.2 Risikoanalyse und Zielvorgaben

Wird für eine Dimension ein *mittlerer* oder *hoher* Schutzbedarf festgestellt, so ist die KI-Anwendung in Hinblick auf jedes der ihr untergeordneten Risikogebiete zu prüfen. Dabei ist es ein wichtiges Ziel des Prüfkatalogs, in der Breite aller denkbaren KI-Anwendungen anwendbar zu sein. Dies verhindert insbesondere quantitative Mindestanforderungen und Schwellwerte zu spezifizieren, die möglicherweise für die eine KI-Anwendung angemessen sind, aber für andere KI-Anwendungen zu schwach oder zu restriktiv wären. Um dennoch zu einem praktikablen Prüfverfahren zu kommen, das sowohl Prüfer*innen als auch Entwickler*innen von KI-Anwendungen eine substanzielle Entwicklung und Prüfung ermöglicht, sieht der Katalog, vergleichbar zum IT-Grundschutz und den Common Criteria[25], vor, zunächst eine für die jeweilige KI-Anwendung spezifische Risikoanalyse durchzuführen und darauf basierend Zielvorgaben (sowie später Kriterien, siehe Abschnitt 2.3.3) zu definieren.

Ziel der Risikoanalyse ist es festzustellen, welche der potenziellen Risiken des Risikogebiets für die spezifische KI-Anwendung relevant sind und beherrscht werden müssen. Dazu werden zunächst die dem Risikogebiet zugehörigen Risiken in Anbetracht des Einsatzkontextes der KI-Anwendung ermittelt und eingeordnet. Die identifizierten Risiken werden anschließend, ähnlich wie bei der Schutzbedarfsanalyse, anhand möglicher resultierender Schäden und Gefährdungen bewertet. Beispielsweise sollte bei einer KI-Anwendung zur Handschriftenerkennung, die eingesetzt wird, um Vorschläge zur Vervollständigung von Wörtern zu geben, eine falsche Ausgabe eher akzeptabel sein als bei einer KI-Anwendung, die Handschriften zur Verifikation von Unterschriften unter Verträgen erkennt.

Anmerkung: Die Auflistung von Risiken im Katalog hat keinen Anspruch auf Vollständigkeit und befreit den*die Entwickler*in oder den*die Prüfer*in nicht von der Pflicht, weitere nicht aufgelistete Risiken zu erkennen und in der Risikoanalyse aufzugreifen.

Basierend auf den als relevant identifizierten Risiken werden Zielvorgaben formuliert. Die Zielvorgaben legen fest, unter welchen Umständen ein vertretbares Restrisiko in Hinblick auf das betrachtete Risikogebiet hergestellt ist. Falls möglich, wird in den Zielvorgaben bereits ein Vorgehen skizziert, durch das das Restrisiko auf ein vertretbares Maß gesenkt werden kann.

2.3.3 Kriterien zur Zielerreichung

Eine wichtige Voraussetzung, um das Erreichen der Zielvorgaben objektiv überprüfen zu können, ist es, nachvollziehbare Kriterien für deren Erreichen zu definieren.

Gerade im Bereich der Entwicklung von KI-Anwendungen gibt es eine Reihe von Metriken (*Key Performance Indicators*, KPIs), die die Güte von Daten oder Modellen beschreiben. Die Wahl einer Metrik zur Überprüfung von Qualitätseigenschaften ist oft abhängig von der KI-Anwendung, jedoch nicht beliebig. Der KI-Prüfkatalog zeigt akzeptierte Metriken sowie Qualitätseigenschaften auf und bietet damit eine Orientierung bei der Wahl anwendungsspezifischer Kriterien. In jedem Fall ist zu begründen, dass diese für die betrachtete KI-Anwendung

---

25 Für weiterführende Informationen und Erläuterungen siehe auch die folgenden Webseiten: https://www.bsi.bund.de/DE/Themen/Unternehmen-und-Organisationen/Standards-und-Zertifizierung/IT-Grundschutz/it-grundschutz_node.html (letzter Aufruf: 01.07.2021) und https://www.bsi.bund.de/DE/Themen/Unternehmen-und-Organisationen/Standards-und-Zertifizierung/Zertifizierung-und-Anerkennung/Zertifizierung-von-Produkten/Zertifizierung-nach-CC/IT-Sicherheiskriterien/CommonCriteria/commoncriteria_node.html (letzter Aufruf: 01.07.2021)

geeignet sind. Für quantitative Metriken sind außerdem anwendungsspezifische Zielintervalle festzulegen, deren Angemessenheit angesichts des Einsatzkontexts dargelegt werden muss. Falls die Zielvorgaben nicht durch quantitative Kriterien abgebildet werden können, wird ferner die Möglichkeit eröffnet, auf rein qualitative Kriterien zurückzugreifen.

Um der Tatsache gerecht zu werden, dass KI-Anwendungen im Betrieb ggf. veränderten Umgebungsbedingungen ausgesetzt sind, werden auch explizit Prüfanforderungen formuliert, die die Betriebsbedingungen der KI-Anwendung betreffen. Solche Zielvorgaben bzw. Kriterien finden sich insbesondere in den Risikogebieten mit dem Titel »Beherrschung der Dynamik« wieder, in denen es unter anderem darum geht, die potenziell durch *Concept Drift* entstehende Dynamik zu beherrschen. Bezüglich der für den Betrieb vorgesehenen Prozesse wird jedoch in erster Linie geprüft, ob deren Umsetzung für die KI-Anwendung ausreichend wäre und nicht notwendigerweise, ob diese Prozesse tatsächlich implementiert sind. Hierdurch wird es möglich, dass ein positives Prüfergebnis für eine KI-Anwendung auch vor Inbetriebnahme erreicht werden kann. Eine Prüfung während des laufenden Betriebs sollte jedoch zusätzlich untersuchen, ob die geforderten Prozesse tatsächlich implementiert und in der realen Einsatzumgebung wirksam sind.

2.3.4 Maßnahmen

Nachdem in einem risikobasierten Ansatz Qualitätskriterien hergeleitet wurden, besteht der nächste Schritt darin, die Erfüllung dieser Kriterien zu belegen. Hierbei erlaubt es der KI-Prüfkatalog, konkrete technische und organisatorische Maßnahmen einzubringen, die relevante Risiken auf ein akzeptables Maß mindern oder die Beherrschung dieser Risiken durch Tests nachweisen. Im Folgenden wird beschrieben, wie der Prüfkatalog bei der Entwicklung einer Absicherungsargumentation für KI-Anwendungen unterstützt. Dazu wird dargestellt, in welchem Umfang und bezüglich welcher Kategorien Maßnahmen eingebracht werden können und wie sich klassische bzw. KI-unspezifische Maßnahmen in das Vorgehen des Katalogs einordnen.

Der KI-Prüfkatalog gibt eine Anleitung zur strukturierten Dokumentation von Maßnahmen entlang des Lebenszyklus einer KI-Anwendung. Im Gegensatz zu klassischen Standards und Prüfverfahren aus funktionaler und IT-Sicherheit stellt sich bei KI-Anwendungen jedoch die besondere Herausforderung, dass sich sowohl die möglichen Gefährdungen, die die Ursache für ein Risiko darstellen, als auch die möglichen Maßnahmen zur Minderung des Risikos aufgrund des breiten Spektrums von KI-Anwendungen nicht vollständig darstellen lassen und sich im Laufe der Zeit weiterentwickeln. Die im Katalog aufgezeigte Vorgehensweise gibt daher eine Orientierung, welche Absicherungs- und Testmaßnahmen für KI-Anwendungen in Betracht gezogen werden können. In jedem Fall sind, falls Methoden aus dem Prüfkatalog ergriffen werden, diese spezifisch auf die KI-Anwendung sowie ihren Einsatzkontext abzustimmen. Ferner ist es nicht zwingend erforderlich und in der Regel auch nicht umsetzbar bzw. verhältnismäßig für eine KI-Anwendung alle im Prüfkatalog angedeuteten Vorkehrungen zu ergreifen. Von Bedeutung ist, dass die ergriffenen und dokumentierten Maßnahmen zusammen betrachtet ausreichen, um relevante Risiken auf ein vertretbares Maß zu senken.

Zudem können auch Maßnahmen ergriffen und dokumentiert werden, die nicht im Prüfkatalog adressiert sind; insbesondere solche, die KI-unspezifisch sind. Denn da der KI-Prüfkatalog als Ergänzung zu bestehenden Prüfverfahren zu verstehen ist, adressiert er vorrangig KI-spezifische Maßnahmen. Maßnahmen zur Abschwächung von Risiken, die auch bei klassischen IT-Systemen bestehen, werden im Prüfkatalog nur dann vereinzelt angeführt, wenn sie einen essenziellen Anteil an der Abschwächung von (zwar KI-unspezifischen, aber durch den Einsatz von KI gesteigerten) Risiken haben. So finden sich in der Dimension Sicherheit im Risikogebiet Integrität und Verfügbarkeit zum Teil Maßnahmen der klassischen IT-Sicherheit wieder, wie etwa der physische Schutz des Speicherorts von Daten sowie die Beschränkung von Abfragemöglichkeiten. Damit wird der Prüfkatalog der Tatsache gerecht, dass Daten im Fall KI-basierter Anwendungen einen noch sensibleren Angriffsvektor darstellen als angesichts klassischer IT-Systeme. Jedoch hat der KI-Prüfkatalog, auch wenn er mitunter KI-unspezifische Maßnahmen aufzeigt, nicht den Anspruch, bestehende Standards, etwa zur IT-Sicherheit, im Kontext von KI-Anwendungen zu ersetzen.

Der Lebenszyklus einer KI-Anwendung eröffnet verschiedenartige Ansatzpunkte zur Abschwächung von Risiken. Orientiert an den Stadien des Lebenszyklus wird bei den Maßnahmen im KI-Prüfkatalog zwischen den folgenden vier Kategorien unterschieden:

1. Daten
2. Entwicklung und Modellbildung der KI-Komponente
3. Einbettung
4. Betrieb der KI-Anwendung

Damit werden sowohl alle Entwicklungsschritte der KI-Anwendung miteinbezogen als auch die Möglichkeit, dass die KI-Anwendung während des Betriebs weiterlernt.

Wird eine Maßnahme ergriffen, so wird je nach Art der Maßnahme, eine **Dokumentation (»Do«)**, ein **Testbericht (»Te«)**, die Beschreibung eines **Prozesses (»Pr«)** oder eine Kombination daraus gefordert. Bei Maßnahmen, die eine Dokumentation (»Do«) vorsehen, ist darauf zu achten, dass in der Dokumentation die Wirksamkeit der Maßnahme ersichtlich und für sachkundige Dritte nachvollziehbar ist. Werden beispielsweise Design-Entscheidungen dokumentiert, so sollte damit einhergehend dargelegt werden, inwiefern diese zur Erfüllung von Qualitätsanforderungen im betrachteten Risikogebiet beitragen. Bei Tests (»Te«) ist es von Bedeutung, dass nicht nur die Testergebnisse, sondern auch das Setting und die Durchführung, d. h. etwa die verwendeten Testdaten, ausführlich beschrieben werden. Darüber hinaus können zur Abschwächung von KI-Risiken auch Prozesse (»Pr«) etabliert werden, welche im Betrieb einzuhalten sind. Falls ein solcher Prozess zur Erfüllung eines Kriteriums erforderlich ist, so sind die vorgesehenen Prozessschritte im Detail zu dokumentieren, auch wenn die KI-Anwendung (noch) nicht im Betrieb ist.

Falls eine Maßnahme zur Abschwächung verschiedenartiger Risiken beiträgt, ist es nicht erforderlich, ein und dieselbe Dokumentation in mehreren Risikogebieten zu wiederholen, sondern es kann auf die in einem Risikogebiet vorhandene Dokumentation verwiesen werden. Gleichermaßen kann der KI-Steckbrief referenziert werden, falls dort bereits Angaben gemacht wurden, die als risikoabschwächende Maßnahme gewertet werden können.

2.3.5 Gesamtbewertung (eines Risikogebiets)

Zum Abschluss eines Risikogebiets wird eine Gesamtbewertung vorgenommen, deren Ziel es ist, unter Würdigung der dokumentierten Maßnahmen nachzuweisen, dass die zuvor festgelegten Qualitätskriterien erfüllt sind.

Hierbei wird ausführlich dargelegt, inwiefern die dokumentierten Maßnahmen, die durchgeführten Tests sowie die für den Betrieb vorgesehenen Prozesse wirksam sind, um die für dieses Risikogebiet formulierten Ziele zu erreichen. Insbesondere erfolgt eine Beurteilung, inwiefern die zuvor definierten, quantitativen und qualitativen Kriterien erfüllt werden. Sofern nicht alle in den Kriterien spezifizierten Anforderungen erfüllt werden, werden die Abweichungen festgehalten. Dies gilt ebenfalls für nur teilerfüllte Anforderungen, bei denen etwa die Kriterien nicht, oder nicht immer, erreicht wurden. Derartige Abweichungen führen nicht automatisch zum Scheitern der Prüfung, sie müssen jedoch in den übergeordneten Bewertungen der Dimensionen berücksichtigt werden.

2.3.6 Zusammenfassende Betrachtung einer Dimension

Dimensionen mit mittlerem oder hohem Schutzbedarf werden, nachdem die KI-Anwendung in Hinblick auf alle Risikogebiete dieser Dimension untersucht wurde, zusammenfassend betrachtet. Hierbei werden insbesondere mögliche Lücken zusammengetragen, die in den Gesamtbewertungen der einzelnen Risikogebiete identifiziert wurden. Die verbleibenden Restrisiken werden angesichts des Schutzbedarfs der Dimension bewertet.

2.3.7 Dimensionsübergreifende Beurteilung der Vertrauenswürdigkeit der KI-Anwendung

Abschließend, nachdem die KI-Anwendung in Hinblick auf die sechs Dimensionen der Vertrauenswürdigkeit untersucht wurde, erfolgt eine dimensionsübergreifende Beurteilung. Ziel dieser Beurteilung ist es, mögliche Lücken in der Abschwächung von Risiken sowie potenzielle Trade-Offs zwischen den Dimensionen der Vertrauenswürdigkeit abzuwägen. Darauf basierend wird ein Urteil über die Vertrauenswürdigkeit der KI-Anwendung gefällt und die Prüfung damit abgeschlossen.

Idealerweise sollten alle Risiken aus den Dimensionen mit mittlerem oder hohem Schutzbedarf auf einem akzeptablen Risikoniveau liegen. Unter Umständen ist dieser Zielzustand jedoch dadurch ausgeschlossen, dass zwischen verschiedenen Dimensionen ein Trade-Off besteht. Beispielsweise könnte eine Erhöhung der Transparenz zur Folge haben, dass es Angreifer*innen ermöglicht wird, basierend auf neu verfügbaren Informationen effektivere und gezielte Angriffe auf die KI-Anwendung zu erstellen. Ein weiteres Beispiel ist, dass einige quantitative Konzepte von Fairness im Widerspruch dazu stehen können, dass die KI-Anwendung eine hohe *Accuracy* hat, falls die realen Testdatensätze aus Sicht des Konzepts »unfair« sind. Dadurch kann sich etwa ein Trade-Off zwischen den Dimensionen Fairness und Verlässlichkeit ergeben.

Somit ist es möglich, dass Kriterien oder Maßnahmen zur Risikoabschwächung in der einen Dimension die Risiken in einer anderen Dimension, die ggf. sogar einen höheren Schutzbedarf hat, steigern. Dies erfordert Abwägung und wird in der dimensionsübergreifenden Beurteilung entsprechend diskutiert.

Der risikobasierte Ansatz des Katalogs ermöglicht es hierbei, verschiedene Anforderungen gegeneinander aufzuwiegen. Insbesondere sind bei der Beurteilung der KI-Anwendung Ausnahmen in Bezug auf bestehende Lücken unter Umständen möglich, falls sich diese in einem vertretbaren Rahmen bewegen und plausibel dargelegt werden kann, dass diese aufgrund von Trade-Offs unvermeidbar sind.

3. KI-Steckbrief (ST)

Der im Folgenden dargestellte KI-Steckbrief bezieht sich auf den Fall, dass der Prüfkatalog zur Prüfung einer KI-Anwendung durch eine externe, dritte Partei verwendet wird. Er hat den Zweck, der prüfenden Instanz vor der eigentlichen Prüfung einen ersten Überblick über die KI-Anwendung hinsichtlich ihrer Funktionalität, des vorgesehenen Einsatzkontextes sowie ihres Aufbaus zu verschaffen. Da der KI-Steckbrief lediglich einen allgemeinen Überblick über die KI-Anwendung bieten und nicht als Basis für eine umfassende Beurteilung des Systems[26] dienen soll, können die Angaben hierbei entsprechend kurzgehalten werden. In den Dokumentationen, die nachfolgend im Rahmen der Prüfung angefordert werden, sollten hingegen möglichst detaillierte Ausführungen und konkrete (technische) Spezifikationen vorgenommen werden.

Der Steckbrief gliedert sich in die Beschreibung der grundlegenden Funktionalität und des vorgesehenen Einsatzkontextes sowie der Struktur der KI-Anwendung.

Grundlegende Funktionalität und vorgesehener Einsatzkontext (FE)

[ST-B-FE-01] Beschreiben Sie die Aufgabe bzw. Funktionalität der KI-Anwendung. Erläutern Sie hierbei auch die folgenden Punkte:

- Welche Aufgabenstellung wird durch die KI-Anwendung gelöst? (Was »macht« sie genau?)
- Welche Eingabedaten sind vorgesehen und welche Form haben sie?
- Was sind die zugehörigen Ausgaben der KI-Anwendung und welche Form haben sie?

26 In der Literatur wurde bereits eine Vielzahl umfänglicher Fragebögen zur Beurteilung der Vertrauenswürdigkeit von KI-Systemen veröffentlicht, darunter auch Fragebögen, die etwa auf spezielle Aspekte der Vertrauenswürdigkeit fokussieren. Ein breites Spektrum wird durch den Fragebogen zur Selbsteinschätzung der HLEG abgedeckt:
High-Level Expert Group on AI (HLEG) (Juli 2020). The Assessment List for Trustworthy Artificial Intelligence (ALTAI). Veröffentlicht von der Europäischen Kommission. https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment (letzter Aufruf: 21.06.2021)
Darüber hinaus gibt es eine ganze Reihe von Beiträgen, mithilfe derer eher technisch orientierte Aspekte von KI-Systemen abgefragt werden können. Für eine strukturierte Übersicht wichtiger Eigenschaften von Datensätzen, etwa zur Genese oder ihrem Verwendungszweck, siehe zum Beispiel:
Gebru et al. (2018). Data Sheets for Datasets. In: Proceedings of the 5th Workshop on Fairness, Accountability, and Transparency in Machine Learning, PLMR 80. https://www.fatml.org/media/documents/datasheets_for_datasets.pdf (letzter Aufruf: 29.06.2021)
Weiterhin bieten:
M. Arnold et al. (2019). »FactSheets: Increasing trust in AI services through supplier's declarations of conformity,« in *IBM Journal of Research and Development*, vol. 63, no. 4/5, pp. 6:1-6:13, 1 July-Sept. 2019. https://ieeexplore.ieee.org/document/8843893. (letzter Aufruf: 29.06.2021); und Mitchell et al. (2019). Model Cards for Model Reporting. In: FAT* '19: Proceedings of the Conference on Fairness, Accountability, and Transparency, S. 220-229; https://doi.org/10.1145/3287560.3287596 (letzter Aufruf: 29.06.2021)
Ansätze zur strukturierten Erfassung wichtiger ML-Modelleigenschaften und -informationen, wobei sie Hersteller-Konformitätserklärungen als Vorbild nehmen.

[ST-B-FE-02] Beschreiben Sie den vorgesehenen Einsatzkontext und die vorgesehene Betriebsumgebung der KI-Anwendung näher. Erläutern Sie hierbei auch die folgenden Punkte:

- Ist die KI-Anwendung in ein Gesamtsystem eingebettet? Ist dies der Fall, so beschreiben sie den Zusammenhang und die Interaktion zwischen KI-Anwendung und Gesamtsystem. Skizzieren Sie hierbei insbesondere die Schnittstellen.
- Inwiefern sind Menschen am Betrieb oder an der Aufsicht über die KI-Anwendung beteiligt?

[ST-B-FE-03] Welche Anforderungen hinsichtlich Regulatorik, Wirtschaftlichkeitsbetrachtung sowie der Vermeidung möglicher materieller und immaterieller Risiken (zum Beispiel Funktionale Sicherheit, IT-Sicherheit, Persönlichkeitsrechte, …) ergeben sich an die KI-Anwendung im Rahmen des vorgesehenen Einsatzkontextes?

[ST-B-FE-04] In welchen Einsatzkontexten ist die KI-Anwendung darüber hinaus denkbar? Und in welchen zu **[ST-B-FE-02]** verwandten Einsatzkontexten bzw. Betriebsumgebungen sollte von der KI-Anwendung abgesehen werden?

[ST-B-FE-05] Gibt es weitere wichtige Informationen zur Funktionalität der KI-Anwendung oder ihrer Betriebsumgebung?

Struktur der KI-Anwendung (ST)

[ST-B-ST-01] Beschreiben Sie den Aufbau der KI-Anwendung. Skizzieren Sie hierfür:

- Eine Auflistung der wichtigsten Komponenten (KI-Komponente, weitere Softwaremodule) sowie die Spezifikation ihrer Funktionalitäten
- Die Architektur der KI-Anwendung sowie das Zusammenspiel der einzelnen Komponenten untereinander

[ST-B-ST-02] Beschreiben Sie die KI-Komponente näher. Machen Sie hierbei die folgenden Angaben:

- Auf welchem ML-Modell bzw. Lernalgorithmus basiert die KI-Anwendung?
- Lernt die KI-Komponente im Betrieb kontinuierlich, in regelmäßigen zeitlichen Abständen oder nach Initiation von Neutrainings weiter?

[ST-B-ST-03] Gibt es weitere wichtige Punkte zur Struktur der KI-Anwendung?

4. Dimension: Fairness (FN)

Beschreibung und Zielsetzung

Als Ausfluss des allgemeinen Gleichbehandlungsgrundsatzes ist sowohl in ethischer als auch in rechtlicher Hinsicht von einer KI-Anwendung die Wahrung des Prinzips der Fairness zu verlangen. Gemeint ist damit das Verbot, gleiche soziale Sachverhalte ungleich oder ungleiche gleichzubehandeln, es sei denn, ein abweichendes Vorgehen wäre sachlich gerechtfertigt. Dies bedeutet insbesondere, dass Individuen nicht aufgrund ihrer Zugehörigkeit zu einer marginalisierten oder benachteiligten Gruppe diskriminiert werden dürfen.[27]

Zum Beispiel darf die KI-Anwendung Personen nicht ungerechtfertigt infolge ihrer Religionszugehörigkeit, ihres Alters oder ihres Geschlechts eine allgemein bevorzugte Ausgabe vorenthalten. So sollte etwa eine KI-Anwendung, die im Recruiting-Prozess darüber entscheidet, ob eine Person zum Bewerbungsgespräch eingeladen wird, nicht ungerechtfertigt Männer bevorzugen. Aber auch wenn es keine allgemein bevorzugte Ausgabe (wie etwa die Einladung zum Bewerbungsgespräch) gibt, kann eine KI-Anwendung diskriminieren. Dies kann sich beispielsweise so gestalten, dass sich die Qualität bzw. Performanz ihrer Ausgaben in Bezug auf bestimmte Personengruppen verringert. Beispielsweise müssen Sprachsteuerungen auch auf Personen mit Akzenten oder Soziolekten reagieren können und individuell anpassbar sein. Darüber hinaus darf Gesichtserkennungssoftware grundsätzlich nicht häufiger Fehler bei Menschen mit einer bestimmten Hautfarbe oder anderen phänotypischen Merkmalen machen.

KI-Anwendungen lernen aus historischen Daten. Diese sind nicht notwendigerweise vorurteilsfrei. Beinhalten die Daten beispielsweise Benachteiligungen von Frauen, so kann das ML-Modell diese Vorurteile übernehmen. Außerdem können in der Datengrundlage bestimmte Gruppen unterrepräsentiert sein. Man spricht dann von Bias. Bias kann ebenfalls zu Entscheidungen führen, die unfair sind. Als abschreckendes Beispiel bekannt geworden ist die fehlerhafte Klassifikation von dunkelhäutigen Menschen als Gorillas durch Google Fotos[28]. Daher müssen repräsentative Trainingsdaten sichergestellt werden. Darüber hinaus kommt als geeignetes Instrument zur Vermeidung von Diskriminierung eine Nachbesserung der Ausgabe des ML-Modells in Betracht.

Um Fairness zu operationalisieren, muss aus technischer Sicht jeweils ein quantifizierbarer Fairnessbegriff entwickelt werden. Dies setzt in einem ersten Schritt voraus, diejenigen Gruppen zu identifizieren, die durch die KI-Anwendung potenziell benachteiligt werden. Dies können gesellschaftliche Minderheiten oder gesellschaftlich benachteiligte Gruppen sein, aber auch Unternehmen oder allgemein juristische Personen, wie es beispielsweise bei der Preisbildung auf digitalen Marktplätzen der Fall ist. Für die anschließende Wahl der Fairness-Definition ist die Unterscheidung von Gruppenfairness und individueller Fairness hervorzuheben. Bei Gruppenfairness wird verlangt, dass die Ergebnisse der KI-Anwendung für alle vorhandenen Gruppen vergleichbar sind, z. B. im Sinne einer gleichen Verteilung der Ausgaben auf die verschiedenen Gruppen oder im Sinne von gleicher »Trefferwahrscheinlichkeit« bzw. Vorhersagequalität in allen Gruppen. Bei individueller Fairness wird die gleiche Behandlung von gleichen Individuen als Maßstab gesetzt.

---

27 Die Darstellung in diesem Abschnitt sowie teils in den folgenden Abschnitten ist stark angelehnt an das Kapitel »3.2 Fairness« des Whitepapers: Poretschkin, M., et al. (2019). Vertrauenswürdiger Einsatz von Künstlicher Intelligenz. Sankt Augustin: Fraunhofer-Institut für Intelligente Analyse und Informationssysteme IAIS. https://www.ki.nrw/wp-content/uploads/2020/03/Whitepaper_KI-Zertifizierung.pdf (letzter Aufruf: 18.06.2021)

28 Kühl, E. (Juli 2015). Gesichtserkennung: Auch selbstlernende Algorithmen müssen begleitet werden. Zeit Online. https://www.zeit.de/digital/internet/2015-07/google-fotos-algorithmus-rassismus/seite-2 (letzter Aufruf: 16.06.2021)

Die Risikogebiete der Dimension Fairness sind gegeben durch:

1. **Fairness**: Dieses Risikogebiet adressiert das Risiko, dass die KI-Anwendung während der Entwicklung unfaires bzw. diskriminierendes Verhalten gegenüber Nutzer*innen oder Betroffenen lernt.

2. **Beherrschung der Dynamik**: Dieses Risikogebiet behandelt Risiken hinsichtlich Fairness, die sich durch Veränderungen der Rahmenbedingungen oder Änderungen im Nutzerverhalten ergeben.

Schutzbedarfsanalyse

Das potenzielle Schadensszenario, mit dem sich die Dimension Fairness in erster Linie auseinandersetzt, ist die Diskriminierung einer bestimmten Personengruppe durch die KI-Anwendung – sei es aufgrund der ethnischen Herkunft der Personen, ihres Geschlechts, Alters, der Religion/Weltanschauung, oder sonstiger Indikatoren. Damit ist diese Dimension insbesondere relevant für sogenannte KI-basierte *Decision Support Systems*, die eine Entscheidung über bzw. Kategorisierung von Personen vornehmen. Beispiele dafür sind die KI-basierte Kreditvergabe, Auswahl von Bewerber*innen und Empfehlung bezüglich medizinischer Behandlung.

Diskriminierung ist ein immaterieller Schaden, der entsteht, wenn die KI-Anwendung Persönlichkeitsrechte verletzt. Darüber hinaus kann Diskriminierung zu weiteren Schäden führen, wie beispielsweise finanzielle Schäden eines Unternehmens durch Rufschädigung. Diese spielen jedoch eine untergeordnete Rolle – der Schutzbedarf wird in dieser Dimension anhand der Auswirkungen auf die Betroffenen ermittelt. Er ergibt sich daraus, in welchem Umfang die Ausgabe der KI-Anwendung die Persönlichkeitsrechte der Betroffenen beeinflusst.

Beispiel: Eine KI-Anwendung, die die Rückfälligkeit von Straftäter*innen prognostiziert und Einfluss in die Strafzumessung findet, birgt ein hohes Schadenspotenzial im Gegensatz zu einer KI-Anwendung, die Vorschläge dafür gibt, welche Personen auf einem in den sozialen Medien hochgeladenen Foto markiert werden könnten.

Der Schutzbedarf wird folgendermaßen kategorisiert:

Hoch	Die KI-Anwendung regelt den Zugang zu essenziell die Persönlichkeit betreffenden Diensten/Aktivitäten oder trifft Entscheidungen, die die Persönlichkeitsrechte weitreichend beeinflussen. **Beispiele**: Vergabe eines Visums, Zulassung zu Schulen/Universitäten, automatisierte Kreditvergabe, Entscheidung über die Art der medizinischen Behandlung
Mittel	Die Ausgabe der KI-Anwendung steht, wenn auch nur im weiteren Sinne, im Zusammenhang mit einer Person. Damit sind nicht nur die KI-Anwendungen gemeint, die eine Entscheidung über eine Person ausgeben bzw. diese kategorisieren, sondern auch jene KI-Anwendungen, die personenspezifische Eingaben verarbeiten (z. B. eine Spracherkennung, die die gesprochenen Sätze des*der Nutzer*in in Text niederschreibt). Die Ausgabe der KI-Anwendung ist weder sensibel noch hat sie maßgebliche Auswirkungen bezüglich der Persönlichkeitsrechte der Betroffenen. **Beispiele:** Empfehlung für Gesichtserkennung auf Fotos in den Sozialen Medien, Klassifikation des Alters einer Person basierend auf einem Foto, Spracherkennungssysteme
Gering	Die KI-Anwendung verarbeitet keine personenbezogenen Daten, die Aufschluss über das Alter, Geschlecht, sexuelle Identität, Religion, Weltanschauung, ethnische Herkunft oder über eine mögliche Behinderung geben. Außerdem ist die Funktion/Ausgabe der KI-Anwendung nicht in einen Prozess oder eine Entscheidung integriert, die die Handlungsoptionen oder die Persönlichkeit einzelner Betroffener betrifft. **Beispiele:** Empfehlung personalisierter Werbung, Prognose von Maschinenausfällen

[FN-S] Dokumentation der Schutzbedarfsanalyse

Anforderung: Do

- Der Schutzbedarf der KI-Anwendung für die Dimension Fairness wird als *gering, mittel* oder *hoch* festgelegt. Die Wahl der Kategorie *gering/mittel/hoch* wird unter Bezugnahme auf die oben angeführte Tabelle ausführlich begründet.

Falls der Schutzbedarf für die Dimension Fairness *gering* ist, so ist keine nähere Betrachtung der einzelnen Risikogebiete erforderlich. Wurde hingegen ein *mittlerer* oder *hoher* Schutzbedarf ermittelt, so muss im Folgenden jedes Risikogebiet genauer untersucht werden.

4.1 Risikogebiet: Fairness (FN)

Das Risikogebiet Fairness soll sicherstellen, dass die Ausgabe der KI-Anwendung keine ungewollte oder ungerechtfertigte Diskriminierung von Personen(gruppen) beinhaltet bzw. verursacht[29]. Typische Gefährdungen dieses Risikogebiets, wie im Allgemeinen Gleichbehandlungsgesetz beschrieben, sind folgende:

Benachteiligung von Personen

- einer bestimmten Nationalität oder ethnischen Herkunft,
- eines bestimmten Geschlechts,
- die einer bestimmten Religion oder Weltanschauung angehören,
- die eine Behinderung haben,
- einer bestimmten Altersgruppe,
- einer bestimmten sexuellen Identität.

Die Relevanz dieser Gefährdungen angesichts der vorliegenden KI-Anwendung ist zu beurteilen. Außerdem sollten weitere Personengruppen in Betracht gezogen und ergänzt werden, falls diese aufgrund des spezifischen Einsatzkontexts oder der Anforderungen der KI-Anwendung diskriminiert werden könnten. Das genaue Vorgehen in der Risikoanalyse für das Risikogebiet Fairness wird im folgenden Abschnitt beschrieben.

4.1.1 Risikoanalyse und Zielvorgaben

[FN-R-FN-RI-01] Identifikation potenziell benachteiligter Gruppen

Anforderung: Do

- Es liegt eine Dokumentation vor, in der mögliche durch die Ausgaben der KI-Anwendung benachteiligte Gruppen bzw. Individuen und deren Charakterisierung mittels der in den Daten vorhandenen sensiblen Merkmale identifiziert werden. Dazu sollen zum einen die o. g. typischen Gefährdungen untersucht und ihre Relevanz für die vorliegende KI-Anwendung beurteilt werden. Zum anderen sind weitere kontext- oder systemspezifische Gefährdungen des Risikogebiets Fairness zu ermitteln und zu analysieren.

[FN-R-FN-RI-02] Festlegung eines geeigneten Fairness-Konzepts

Anforderung: Do

- Es liegt eine Dokumentation vor, in der ausführlich beschrieben wird, was Fairness im spezifischen Anwendungskontext der KI-Anwendung bedeutet. Dabei soll insbesondere dargelegt werden, welche Arten von Diskriminierung (d. h. anhand welcher sensiblen Attribute) akzeptabel oder sogar zweckmäßig, und welche Arten von Diskriminierung ungerechtfertigt oder zumindest unerwünscht sind.
  Beispiel: Wenn sich die Höhe eines Versicherungstarifs u. a. nach dem Eintrittsalter einer Person richtet, könnte dies zweckmäßig sein, da das Alter mit dem potenziellen finanziellen Aufwand der Versicherung für diese Person korrelieren könnte. Jedoch ist Diskriminierung aufgrund des Geschlechts beim Tarif einer Kfz-Versicherung ungerechtfertigt[30].

---

29 Ein ähnlicher Ansatz zu dem in diesem Risikogebiet dargestellten Vorgehen für die Entwicklung einer Absicherungsargumentation für die Fairness von algorithmischer Entscheidungsfindung wird auch im folgenden Paper skizziert. Hierbei werden die Konzepte des Acceptance Test-Driven Development (ATDD) und der Assurance Cases kombiniert: Hauer, M. P., Adler, R., & Zweig, K. (2021). Assuring Fairness of Algorithmic Decision Making. 2021 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW). https://doi.org/10.1109/icstw52544.2021.00029 (letzter Aufruf: 30.06.2021)

30 Sommer, M. (November 2012). Der Lady-Tarif hat ausgedient. Zeit Online. https://www.zeit.de/auto/2012-11/autoversicherung-unisex (letzter Aufruf: 16.06.2021)

- Außerdem wird die Konfliktfreiheit der gewählten anwendungsspezifischen Abgrenzung zwischen akzeptierter und unerwünschter Diskriminierung mit geltendem Recht dokumentiert (gerechtfertigte/ungerechtfertigte Diskriminierung). Insbesondere ist hierbei auf die Kompatibilität mit dem Grundsatz der allgemeinen Gleichbehandlung einzugehen.
- **Zielvorgaben:** Für die in **[FN-R-FN-RI-01]** als relevant identifizierten Personengruppen soll das Eintreten sowohl unerwünschter als auch ungerechtfertigter Diskriminierung verhindert bzw. behoben werden.

4.1.2 Kriterien zur Zielerreichung

Basierend auf den identifizierten Gefährdungen des Risikogebiets Fairness sollen entsprechende Absicherungsmaßnahmen getroffen werden. Um jedoch in der abschließenden Bewertung der Maßnahmen objektiv überprüfen zu können, ob die vorhandenen Risiken erfolgreich mitigiert wurden, muss die in **[FN-R-FN-RI-02]** beschriebene Zielvorgabe zunächst in quantitative Kriterien übersetzt werden. Dabei wird zwischen Kriterien, die die Fairness in Bezug auf die Ausgaben der KI-Anwendung über eine sogenannte Fairness-Metrik quantifizieren und Kriterien, die den Bias in den Trainingsdaten quantifizieren, unterschieden. Bei der Wahl der Fairness-Metrik ist insbesondere zu beachten und ausführlich zu begründen, dass sie im Einklang mit der in **[FN-R-FN-RI-02]** beschriebenen Bedeutung von Fairness im Kontext der KI-Anwendung steht.

[FN-R-FN-KR-01] Quantifizierung von Fairness im Output

Anforderung: Do

- Es liegt eine Dokumentation vor, in der die formale Definition der in **[FN-R-FN-RI-01]** erfassten, möglicherweise benachteiligten Gruppen in Form einer geeigneten Kategorisierung über Merkmale und Merkmalskombinationen festgehalten wird.
- Außerdem wird beschrieben, welche formalen Fairness-Definition(en) ausgewählt und welche quantitative(n) Fairness-Metrik(en) daraus abgeleitet wurde(n), mit deren Hilfe die Fairness der KI-Anwendung beurteilt werden soll.
  - Zulässige Definitionen von Fairness (und zugehörige Metriken) sind in einer separaten Anlage zu finden. Wird keine der dort genannten Definitionen verwendet, so muss begründet werden, weshalb nicht. In diesem Fall muss die eigene, stattdessen verwendete Definition bzw. das eigene Kriterium in den Anwendungskontext eingeordnet und die Wahl umfassend begründet werden.
  - Aus der Dokumentation muss ersichtlich sein, wie ggf. mit sich widersprechenden Fairness-Definitionen umgegangen wird.
- Für die gewählten Metriken werden zu erreichende Zielintervalle festgelegt. Hierbei wird ausführlich begründet, dass die gewählte(n) Definition(en), Metrik(en) und Zielintervalle konsistent sind mit den Zielvorgaben sind.

Anlage: Mögliche Fairness-Definitionen[31] sind die folgenden:

- *Group Fairness (Statistical/Demographical Parity, Equal Acceptance Rate, Benchmarking)*
- *Conditional Statistical Parity*
- *Predictive Parity (Outcome Test)*
- *False Positive Error Rate Balance (Predictive Equality)*
- *False Negative Error Rate Balance (Equal Opportunity)*
- *Equalized Odds (Conditional Procedure Accuracy Equality, Disparate Mistreatment)*
- *Conditional Use Accuracy Equality*
- *Overall Accuracy Equality*
- *Treatment Equality*

---

31 Für einen Überblick zu gängigen Fairness-Metriken sowie möglichen Vor- und Nachteilen der Metriken siehe auch: Verma, S., Rubin, J. (2018). Fairness Definitions Explained. 2018 ACM/IEEE International Workshop on Software Fairness. https://doi.org/10.1145/3194770.3194776 (letzter Aufruf: 30.06.2021)

- *Test Fairness (Calibration, Matching Conditional Frequencies)*
- *Well Calibration*
- *Balance for Positive Class*
- *Balance for Negative Class*
- *Causal Discrimination*
- *Fairness through Unawareness*
- *Fairness through Awareness (Individual Fairness[32])*
- *Counterfactual Fairness*
- *No Unresolved Discrimination*
- *No Proxy Discrimination*
- *Fair Inference*

Die obigen Definitionen beziehen sich auf den Fall binärer Klassifikation. Sie lassen sich auf Klassifikation mit k-Klassen ausweiten und unter Umständen auch auf Regression anwenden. Aus den gelisteten Fairness-Definitionen lassen sich quantitative Maße für Fairness ableiten, etwa indem man den Absolutbetrag der Differenz der Größen aus der Definition nimmt.

Wird keines der genannten Maße verwendet, so muss begründet werden, weshalb nicht. In diesem Fall muss die eigene, stattdessen verwendete Definition bzw. das eigene Maß in den Anwendungskontext eingeordnet und umfassend erklärt werden.

[FN-R-FN-KR-02] Quantifizierung von Fairness in den Trainingsdaten

Anforderung: Do

- Es liegt eine Dokumentation vor, in der die Wahl eines oder mehrerer quantitativer Maße zur Beurteilung von Bias in den Trainingsdaten (oder ggf. das Absehen von näherer Untersuchung der Trainingsdaten) beschrieben und begründet wird.
- Für die Maße bezüglich der Trainingsdaten werden geeignete Zielintervalle definiert. Dabei ist die Angemessenheit der Zielintervalle in Bezug auf den Einsatzkontext der KI-Anwendung zu begründen.

4.1.3 Maßnahmen

4.1.3.1 Daten

[FN-R-FN-MA-01] Überprüfung der Daten auf Bias-Freiheit

Anforderung: Do

- Es liegt eine Dokumentation vor, welche die Überprüfung der Daten auf Bias-Freiheit (insbesondere bzgl. der möglicherweise benachteiligten Gruppen) festhält. Dabei werden die geprüften Daten sowie die gemäß **[FN-R-FN-KR-02]** gewählten Maße und erreichten Zielintervalle angegeben.

[FN-R-FN-MA-02] Faire Datenvorverarbeitung

Anforderung: Do

- Es liegt eine Dokumentation vor, aus der hervorgeht, welche Schlüsse man aus **[FN-R-FN-MA-01]** zieht und wie man ggf. die Daten im Sinne einer fairen Vorverarbeitung (faires *Pre-Processing*), beispielsweise durch
  - *Data Massaging,*
  - *Uniform/Preferential Sampling,*

32 Dwork, C., et al. (Januar 2012). Fairness Through Awareness. In: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference 2012, pp. 214-226. https://doi.org/10.1145/2090236.2090255 (letzter Aufruf am 30.06.2021)

  - *Reweighing*[33],
  - Faire Datenrepräsentationen[34],

  aufbereitet.
- In der Dokumentation muss auch festgehalten sein, warum die ergriffenen Maßnahmen in Bezug auf die gewählte Fairness-Metrik sinnvoll sind, und inwiefern sie dazu beitragen, die Fairness der KI-Anwendung zu gewährleisten oder zu verbessern.
- Falls keine solche Vorverarbeitung stattfindet, so muss auch dies begründet werden.

4.1.3.2 KI-Komponente

[FN-R-FN-MA-03] Faire Modellbildung

Anforderung: Do

- Es liegt eine Dokumentation vor, die Angaben zum verwendeten Modell macht und beschreibt, wie durch das implementierte Lernverfahren[35] und insbesondere die gewählten Verlustfunktion(en)[36] die Fairness der KI-Anwendung gefördert wird.

[FN-R-FN-MA-04] Faire Adaption und Nachverarbeitung

Anforderung: Do

- Es liegt eine Dokumentation vor, die beschreibt, welche Maßnahmen ergriffen werden, falls die Ergebnisse des ML-Modells im Lernprozess unfair werden (Faires *In-Processing*, *Optimization at Training Time*). Zur Beurteilung werden dabei die Zielintervalle gemäß **[FN-R-FN-KR-01]** zugrunde gelegt. Die Maßnahmen müssen begründet sein.
- Es liegt eine Dokumentation vor, die beschreibt, welche Maßnahmen ergriffen werden, falls die Ergebnisse nach dem Training unfair werden (Faires *Post-Processing*[37]). Zur Beurteilung werden die Zielintervalle gemäß **[FN-R-FN-KR-01]** zugrunde gelegt. Diese Maßnahmen müssen begründet sein.
- Falls keine solchen Maßnahmen ergriffen werden, so muss auch dies begründet werden.

---

33 Für *Data Massaging*, *Uniform/Preferential Sampling* und *Reweighing* siehe auch: Kamiran, F., Calders, T. (Dezember 2011). Data preprocessing techniques for classification without discrimination. Springer, Knowledge and Information Systems 33, 1–33 (2012). https://doi.org/10.1007/s10115-011-0463-8 (letzter Aufruf: 30.06.2021)

34 Siehe auch: Zemel, R., Wu, Y., Swersky, K., Pitassi, T. und Dwork, C. (2013). Learning fair representations. In Proceedings of the 30th International Conference on Machine Learning, PMLR 28(3):325-333, 2013. http://proceedings.mlr.press/v28/zemel13.pdf (letzter Aufruf: 30.06.2021) sowie Lahoti, P., Gummadi, K. und Weikum, G. (2019). ifair: Learning individually fair data representations for algorithmic decision making. In 35th IEEE International Conference on Data Engineering, 2019. https://doi.org/10.1109/ICDE.2019.00121 (letzter Aufruf: 30.06.2021)

35 Für eine mögliche Methode zur fairen Modellbildung siehe beispielsweise: Zhang, B., Lemoine, B. und Mitchell, M. (2018). Mitigating Unwanted Biases with Adversarial Learning. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES ,18). Association for Computing Machinery, New York, NY, USA, 335–340. https://doi.org/10.1145/3278721.3278779 (letzter Aufruf: 30.06.2021)

36 Für mögliche Modifikationen der Verlustfunktion zur fairen Modellbildung siehe beispielsweise: Zafar, M., Valera, I., Gomez Rodriguez, M. und Gummadi, K. (2017). Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. In Proceedings of the 26th International Conference on World Wide Web (WWW ,17). International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 1171–1180. https://doi.org/10.1145/3038912.3052660 (letzter Aufruf: 30.06.2021)

37 Siehe hierzu auch: Hardt, M., Price, E. und Srebro, N. (2016). Equality of opportunity in supervised learning. In Advances in Neural Information Processing Systems 29, 2016. https://papers.nips.cc/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf (letzter Aufruf: 30.06.2021) sowie F. Kamiran, A. Karim and X. Zhang (2012), »Decision Theory for Discrimination-Aware Classification,« 2012 IEEE 12th International Conference on Data Mining, pp. 924-929, https://doi.org/10.1109/ICDM.2012.45 (letzter Aufruf: 30.06.2021)

[FN-R-FN-MA-05] Tests der KI-Komponente auf ungesehenen Daten

Anforderungen: Do | Te

- Es werden Tests der KI-Komponente auf im Training ungesehenen Daten durchgeführt und deren Bedeutung für die Fairness der KI-Anwendung dokumentiert. Die erreichten Zielintervalle müssen ebenfalls angegeben werden.

4.1.3.3 Einbettung

[FN-R-FN-MA-06] Faire Weiterverarbeitung

Anforderung: Do

- Es liegt eine Dokumentation vor, aus der hervorgeht, welche möglicherweise Fairness-relevanten Verarbeitungsschritte durch Komponenten der Einbettung die Ausgaben der KI-Komponente durchlaufen.
- Es wird beschrieben, wie sichergestellt wird, dass diese Weiterverarbeitung fair ist. Dabei wird insbesondere dargestellt, wie in **[FN-R-FN-MA-05]** festgestellte Schwächen adressiert werden.

[FN-R-FN-MA-07] Tests der KI-Anwendung

Anforderungen: Do | Te

- Es werden umfangreiche Tests der KI-Anwendung hinsichtlich Fairness durchgeführt und dokumentiert. Die im Test verwendeten Daten müssen beschrieben und deren Auswahl begründet werden. Die erreichten Zielintervalle müssen ebenfalls angegeben werden. Insbesondere werden im Rahmen der Tests Fairness-relevante Verarbeitungsschritte, die durch Komponenten der Einbettung geleistet werden, überprüft.

4.1.3.4 Maßnahmen für den Betrieb

[FN-R-FN-MA-08] Kontrolle der Ausgaben im Betrieb

Anforderungen: Do | Pr | Te

- Es liegt eine Dokumentation vor, in der beschrieben wird, wie die Fairness der Ausgaben der KI-Anwendung im Produktivbetrieb überwacht wird.

4.1.4 Gesamtbewertung

[FN-R-FN-BW] Gesamtbewertung

Anforderung: Do

- Es gibt eine Dokumentation, in der bestätigt wird, dass die quantitativen Kriterien erfüllt werden.
- Des Weiteren muss eine Beurteilung erfolgen, inwiefern die nicht-quantitativen Kriterien durch die für den Betrieb getroffenen Maßnahmen erreicht werden.
- Sofern nicht alle in **[FN-R-FN-KR-01]** und **[FN-R-FN-KR-02]** spezifizierten Anforderungen erfüllt werden, sind die Abweichungen zu dokumentieren. Dies gilt ebenfalls für nur teilerfüllte Anforderungen, bei denen etwa die Kriterien nicht, oder nicht immer, erreicht wurden.

4.2 Risikogebiet: Beherrschung der Dynamik (BD)

Das Risikogebiet Beherrschung der Dynamik stellt sicher, dass die Fairness der KI-Anwendung auch während des Betriebs aufrechterhalten wird. Insbesondere ergeben sich Herausforderungen bei KI-Anwendungen, die auf neu einkommenden Daten weiterlernen. Ferner können Veränderungen der Rahmenbedingungen, wie beispielsweise Gesetzesänderungen, auch nach Inbetriebnahme der KI-Anwendung Maßnahmen erfordern. Grundlegend bestehen für dieses Risikogebiet die folgenden zwei Gefährdungen:

1. ***Model Drift*:** Das Modell lernt Diskriminierung durch während des Betriebs neu einkommende Trainingsdaten.
   Beispiel: Die KI-Anwendung wird in regelmäßigen Abständen auf Daten nachtrainiert, die etwa durch Crowdsourcing oder Nutzereingaben gelabelt werden. Diese könnten aufgrund von Trends oder gesellschaftlichen Ereignissen, im Gegensatz zu den ursprünglichen Trainingsdaten, einem Bias unterliegen.

2. ***Concept Drift*:** Veränderte äußere Bedingungen stellen neue Anforderungen an eine Fairness-Definition.
   Beispiel: Eine Gesetzesänderung besagt, dass sich die Höhe des Versicherungstarifs nicht mehr aufgrund des Geschlechts unterscheiden darf.

4.2.1 Risikoanalyse und Zielvorgaben

[FN-R-BD-RI-01] Dokumentation der Risikoanalyse

Anforderung: Do

- **Risikoanalyse:** Es liegt eine Dokumentation vor, in der beschrieben wird, ob und in welchem Umfang die KI-Anwendung während des Betriebs weiterlernt und welche Risiken sich daraus für die Fairness der KI-Anwendung ergeben.
- **Zielvorgaben:** Im Falle eines Weiterlernens im Betrieb wird dargelegt, welche Anforderungen an die neu einkommenden Daten, aus denen die KI-Anwendung weiterlernt, gestellt werden und welche Prozesse oder Mechanismen zur Kontrolle der neu einkommenden Daten bzgl. deren Fairness bestehen. Es wird außerdem beschrieben, wie sichergestellt werden soll, dass die KI-Anwendung während des Betriebs fair bleibt.

4.2.2 Kriterien zur Zielerreichung

[FN-R-BD-KR-01] Bewahrung von Fairness der KI-Anwendung

Anforderung: Do

- Es werden angemessene, anwendungsspezifische Prüfintervalle zur Beurteilung der Fairness in den Ausgaben der KI-Anwendung festgelegt. Die Beurteilung erfolgt gemäß der in **[FN-R-FN-KR-01]** gewählten Metriken und Zielintervalle. Die Wahl der Prüfintervalle wird dokumentiert und begründet.

[FN-R-BD-KR-02] Bewahrung von Fairness in den Trainingsdaten

Anforderung: Do

- Es werden angemessene, anwendungsabhängige Prüfintervalle zur Beurteilung der Fairness in den Trainingsdaten gemäß der in **[FN-R-FN-KR-02]** gewählten Metriken und Zielintervalle festgelegt, dokumentiert und begründet.

4.2.3 Maßnahmen

4.2.3.1 Daten

[FN-R-BD-MA-01] Überwachung der Trainingsdaten

Anforderungen: Do | Pr

- Es gibt einen Prozess, der die sich durch einkommende Daten neu bildenden Trainingsdaten vor ihrer Verwendung mittels der in **[FN-R-FN-KR-02]** gewählten Maße sowie in den in **[FN-R-BD-KR-02]** festgelegten Prüfintervallen auf ihre Bias-Freiheit überprüft. Der Prozess wird in einer Dokumentation beschrieben.

4.2.3.2 KI-Komponente

Für diese Kategorie sind keine Maßnahmen vorgesehen.

4.2.3.3 Einbettung

Für diese Kategorie sind keine Maßnahmen vorgesehen.

4.2.3.4 Maßnahmen für den Betrieb

[FN-R-BD-MA-02] Anwendungsüberwachung

Anforderungen: Do | Pr | Te

- Es gibt einen dokumentierten Prozess, der sicherstellt, dass die Ausgaben der KI-Anwendung gemäß der in **[FN-R-BD-KR-02]** spezifizierten Prüfintervalle auf Konformität mit den gewählten Fairness-Definitionen aus **[FN-R-FN-KR-01]** hin überprüft werden. Hierfür muss auch im laufenden Betrieb bzw. zu Test-Zeiten bekannt sein, welche sensiblen Merkmale es gibt bzw. welche Gruppen potenziell benachteiligt sind.

[FN-R-BD-MA-03] Anwendungsverbesserung

Anforderungen: Do | Pr

- Es gibt einen Prozess, durch den bei Feststellung von unfairem Verhalten der KI-Anwendung oder bei Identifikation von neuartiger Diskriminierung das ML-Modell und die KI-Anwendung verbessert werden.
- Die Anwendungsverbesserungen müssen in einer Dokumentation festgehalten werden und stets nachvollziehbar sein. Bei dieser Verbesserung darf keine Überanpassung in dem Sinne erfolgen, dass andere Betroffene benachteiligt werden. Aus obiger Dokumentation muss daher auch hervorgehen, wie sichergestellt wird, dass die Verbesserungen fair sind und keine neue, ungerechtfertigte Diskriminierung erzeugen.

[FN-R-BD-MA-04] Überwachung äußerer Faktoren

Anforderungen: Do | Pr

- Es gibt einen Prozess zur Überwachung externer Faktoren, die für die Fairness der KI-Anwendung relevant sind. Beispielsweise können neuartige Formen ungerechtfertigter Diskriminierung auftreten und sich in den Daten widerspiegeln, oder es könnten Änderungen der Gesetzeslage beschlossen werden. Hierzu werden Personen mit der Aufgabe betraut, die Entwicklung der äußeren Umstände zu beobachten und zu beurteilen und, falls für notwendig befunden, Änderungsvorhaben an der KI-Anwendung einzuleiten.

4.2.4 Gesamtbewertung

[FN-R-BD-BW] Gesamtbewertung

Anforderung: Do

- Es wird dargelegt, dass ein Prozess zur regelmäßigen Überprüfung der KI-Anwendung sowie der Datenbasis aufgesetzt wurde, der die Kriterien **[FN-R-BD-KR-01]** und **[FN-R-BD-KR-02]** erfüllt.
- Sofern nicht alle in **[FN-R-BD-KR-01]** und **[FN-R-BD-KR-02]** spezifizierten Anforderungen erfüllt werden, sind die Abweichungen zu dokumentieren. Dies gilt ebenfalls für nur teilerfüllte Anforderungen, bei denen etwa die Kriterien nicht, oder nicht immer, erreicht wurden.

Zusammenfassende Betrachtung

[FN-Z] Zusammenfassende Betrachtung der Dimension

Anforderung: Do

- Falls für diese Dimension ein mittlerer oder hoher Schutzbedarf besteht, ist eine Dokumentation über die verbleibenden Restrisiken zu erstellen. Zunächst werden die Restrisiken aus den verschiedenen Risikogebieten dieser Dimension zusammengefasst. Anschließend wird unter Berücksichtigung des Schutzbedarfs analysiert, ob die identifizierten Restrisiken insgesamt als vernachlässigbar, nicht vernachlässigbar (aber vertretbar) oder unvertretbar zu bewerten sind. Das Ergebnis der Analyse ist zu erläutern.
- Falls potenziell negative Auswirkungen von Risiken oder Maßnahmen dieser Dimension auf andere Dimensionen wie etwa Verlässlichkeit festgestellt wurden, sind diese zu dokumentieren.
- Es wird ein Fazit über die Dimension gezogen, welches insbesondere die Bewertung der Restrisiken enthält.

5. Dimension: Autonomie und Kontrolle (AK)

Beschreibung und Zielsetzung

Autonomie kann sowohl in Bezug auf die Menschen betrachtet werden, die eine KI-Anwendung nutzen oder dadurch potenziell beeinflusst werden, als auch in Bezug auf die KI-Anwendung selbst. Aus Sicht der Informatik ist Autonomie eine Anforderung an ein System, um in einer komplexen Umgebung zu operieren, die Unsicherheit aufweist[38]. Autonomie basiert insbesondere auf einer intrinsischen Motivation des Systems, die sein Verhalten in unsicheren Situationen lenkt. Eine mögliche Herangehensweise an komplexe Problemstellungen, die einen gewissen Grad an Autonomie erfordern, da Automation nicht möglich oder zielführend erscheint, sind ML-Verfahren, die selbständig Modelle bzw. Entscheidungsregeln aus Daten lernen. Gleichzeitig eröffnen KI-Komponenten, die motivationale Prozesse (teil-)autonomer Anwendungen steuern, ein Spannungsfeld zur Autonomie von Nutzer*innen und Betroffenen[39]. So sind KI-Anwendungen schlimmstenfalls etwa in der Lage, Menschen zu überwachen und zu kontrollieren, sie zu bestimmten Handlungen hinzulenken, sie zu täuschen und zu manipulieren. Dies kann beispielsweise geschehen durch Auswahl und Bereitstellung von Informationen gegenüber den Nutzer*innen bzw. durch intelligente, Nutzer-spezifische Interaktion. Infolge unangemessener Autonomie können KI-Anwendungen in die persönlichen Grundrechte des*der Einzelnen, d. h. in seine*ihre Freiheitsrechte (Art. 2 Grundgesetz) und gegebenenfalls auch in seine*ihre Würde (Art. 1 Grundgesetz) eingreifen. Dabei geht es um das Recht des*der Einzelnen, informiert und selbstbestimmt Entscheidungen zu treffen, die seine*ihre eigene Rechtsposition betreffen. Grundrechte schützen auch die freie Meinungsbildung (Art. 5 Grundgesetz) und die Ausübung von demokratischen Rechten.

Das Spannungsfeld zwischen der Autonomie der KI-Anwendung und der Autonomie von Nutzer*innen und Betroffenen muss entsprechend kontrolliert werden. Insbesondere ergibt sich die Anforderung, im Einsatz einer KI-Anwendung den Vorrang menschlichen Handelns zu wahren. Der Vorrang des menschlichen Handelns beinhaltet vor allem auch einen gestalterischen Auftrag für eine angemessene und verantwortungsvolle Rollenverteilung zwischen Mensch und KI-Anwendung. Schon beim Entwurf einer Anwendung sollte durch das partizipative Einbeziehen von Nutzer*innen und Expert*innen für eine verantwortungsvolle Gestaltung gesorgt werden. Der Autonomiegrad der KI-Anwendung muss dem Anwendungskontext angemessen sein und die für Nutzer*innen notwendigen Eingriffs- und Aufsichtsmöglichkeiten gewährleisten. Menschliche Aufsicht im Kontext von Künstlicher Intelligenz bedeutet, dass eine KI-Anwendung während ihres Betriebs durch Menschen überwacht wird und Teilfunktionen oder sogar die ganze KI-Anwendung abgeschaltet werden können. Hierbei ist schon im Vorfeld darauf zu achten, ob eine Abschaltung nicht durch eine Abhängigkeit von der KI-Anwendung faktisch unmöglich gemacht wird. Dies würde das Gebot der menschlichen Aufsicht aushebeln.

---

38 Die Definition des Autonomiebegriffs ist angelehnt an die Darstellung in: Abbass H.A., Scholz J., Reid D.J. (2018). Foundations of Trusted Autonomy: An Introduction. In: Studies in Systems, Decision and Control, vol 117. Springer, Cham. https://doi.org/10.1007/978-3-319-64816-3_1 (letzter Aufruf: 21.06.2021)

39 Für eine tiefgehende Betrachtung des Spannungsfelds zwischen der Autonomie des Menschen und der KI siehe auch: von Braun, J., Archer, M. S., Reichberg, G. M., Sánchez Sorondo, M. (2021). AI, Robotics, and Humanity: Opportunities, Risks, and Implications for Ethics and Policy. Springer, Cham. https://doi.org/10.1007/978-3-030-54173-6 (letzter Aufruf: 21.06.2021)

Darüber hinaus verlangt der Vorrang menschlichen Handelns, dass der*die Einzelne umfassend informiert und befähigt wird, kompetente Entscheidungen treffen. Insbesondere ist im Sinne der Wahrung der Nutzerautonomie dafür zu sorgen, dass eine etwaige Delegation von Entscheidungskompetenzen an eine KI-Anwendung explizit definiert und gewollt ist und nicht im Verborgenen stattfindet. Nutzer*innen und Betroffene müssen über Funktionsweise und mögliche Risiken von KI-Anwendungen aufgeklärt sein, wie auch über ihre Rechte und Beschwerdemöglichkeiten.

Bezüglich dieser Dimension gibt es einer Reihe weiterführender Fragestellungen, etwa, wie damit umgegangen werden soll, dass der Einsatz von KI-Anwendungen am Arbeitsplatz für den*die Einzelnen auch die Dequalifizierung seiner*ihrer Tätigkeit bedeuten kann. Eine andere Problematik ist die, dass im Zuge des Einsatzes einer KI-Anwendung die Verantwortung für eine Entscheidung unter Umständen nicht mehr klar einem Menschen zugeordnet werden. Dies betrifft nicht nur persönliche Entscheidungen, sondern auch Entscheidungsprozesse und Verantwortlichkeiten in Behörden und Unternehmen und kann in einem Schadensfall zu juristischen Problemen führen. Fragestellungen dieser Art sind diesem KI-Prüfkatalog übergeordnet und werden an dieser Stelle nicht weiter behandelt.

Die Risikogebiete für die Dimension Autonomie und Kontrolle sind die folgenden::

1. **Angemessene und verantwortungsvolle Gestaltung der Aufgabenverteilung zwischen Mensch und KI-Anwendung:** Dieses Risikogebiet behandelt Risiken, die sich aus Einschränkungen der Nutzerautonomie bzw. unangemessener Autonomie der KI-Anwendung ergeben.

2. **Sicherstellung der Informiertheit und Befähigung von Nutzer*innen und Betroffenen:** Dieses Risikogebiet adressiert Risiken, die dadurch entstehen, dass Nutzer*innen und Betroffene unzureichend über die KI-Anwendung, deren Nutzung sowie die damit verbundenen Risiken aufgeklärt werden.

Schutzbedarfsanalyse

Das potenzielle Schadensszenario, mit dem sich die Dimension Autonomie und Kontrolle auseinandersetzt, ist die potenzielle Einschränkung der Wahrnehmungs- oder Handlungsfähigkeit von Nutzer*innen oder Betroffenen der KI-Anwendung.

Die Aufgabenverteilung und Interaktionsmöglichkeiten zwischen KI-Anwendung und Nutzer*in müssen daher verantwortungsvoll, klar und transparent geregelt sein. Dabei muss dem*der Nutzer*in in angemessenem Umfang die Möglichkeit zur Steuerung der Anwendung verliehen werden. Generell gilt, dass Nutzer*innen und Betroffene mit den möglichen Risiken bzgl. einer eventuellen Beeinträchtigung ihrer Wahrnehmungsfähigkeit oder Handlungsfreiheit, mit ihren Rechten, Pflichten und Eingriffs- sowie Beschwerdemöglichkeiten vertraut gemacht werden müssen. In Bezug auf Künstliche Intelligenz ist insbesondere darzulegen, inwiefern individuelle Nutzer*innen übermäßiges Vertrauen in die KI-Anwendung entwickeln, emotionale Bindungen aufbauen oder in ihrer Entscheidungsfindung unzulässig beeinträchtigt bzw. gelenkt werden könnten.[40]

Die Einschränkung der Wahrnehmungs- oder Handlungsfähigkeit ist ein immaterieller Schaden, der grundlegende anerkannte ethische und rechtliche Werte betrifft. Darunter fallen die Freiheit des Einzelnen, selbstbestimmt Entscheidungen zu treffen, sowie die Freiheit, die Ziele des eigenen Verhaltens ebenso wie die Wahl der Mittel zur Erreichung dieser Ziele zu bestimmen. Darüber hinaus kann eine Einschränkung der Wahrnehmungs- oder Handlungsfähigkeit zu weiteren Schäden führen, wie beispielsweise Sach- oder Personenschäden, die durch autonom agierende Fahrzeuge oder Roboter hervorgerufen werden. Diese spielen jedoch eine untergeordnete Rolle – der Schutzbedarf wird in dieser Dimension anhand der Auswirkungen auf die Nutzer*innen und

40 Die Darstellung in diesem Abschnitt ist stark angelehnt an das Kapitel »3.1 Autonomie und Kontrolle« des Whitepapers: Poretschkin, M., et al. (2019). Vertrauenswürdiger Einsatz von Künstlicher Intelligenz. Sankt Augustin: Fraunhofer-Institut für Intelligente Analyse und Informationssysteme IAIS. https://www.ki.nrw/wp-content/uploads/2020/03/Whitepaper_KI-Zertifizierung.pdf (letzter Aufruf: 18.06.2021)

Betroffenen ermittelt. Genauer gesagt, ergibt sich dieser daraus, in welchem Umfang die KI-Anwendung die Wahrnehmungs- und Handlungsfähigkeit der Nutzer*innen und Betroffenen beeinflusst. Insbesondere haben KI-Anwendungen, die per se keine oder geringe Eingriffsmöglichkeiten für Nutzer*innen bieten, einen hohen Schutzbedarf.

Beispiel: Eine KI-Anwendung, die automatisiert Preise für Zugtickets festlegt, birgt ein höheres Schadenspotenzial als eine KI-Anwendung, die auf Basis von Nachfrageprognosen einen Ticketpreis vorschlägt.

Der Schutzbedarf wird folgendermaßen kategorisiert:

Hoch	Die KI-Anwendung hat einen hohen Schutzbedarf bezüglich dieser Dimension, falls sie ▪ die Wahrnehmung oder Handlungen von Nutzer*innen oder Betroffenen über lange Zeiträume hinweg oder unter nicht vertretbarem Risiko stark beeinflusst. ▪ die Wahrnehmungs- oder Handlungsfähigkeit von Nutzer*innen oder Betroffenen einschränkt. **Beispiele**: Eine KI-Anwendung, die die Stimme von Pflegepersonal so moduliert, dass demente Patient*innen glauben, sie würden mit einem*einer nahestehenden Angehörigen sprechen. Ein autonom fahrendes Fahrzeug, das auch Menschen transportiert. Eine KI-Anwendung, die den Zugang zu Bildung regelt, indem sie z. B. über die Zulassung an einer Universität entscheidet und die Betroffenen nicht über die Verwendung der KI-Anwendung informiert.
Mittel	Die KI-Anwendung kann die Wahrnehmung oder Handlungen von Nutzer*innen oder Betroffenen vorübergehend und nur unter vertretbarem Risiko stark beeinflussen. **Beispiele:** Eine KI-Anwendung, die das Verhalten des*der Nutzer*in in Bezug auf Fitness, Gesundheit und Ernährung erfasst und mit Vorschlägen und Vorgaben steuert. Eine KI-Anwendung in Form einer Puppe, die durch Sprach-Ein- und Ausgabe sowie Mimik und Bewegung menschliche Interaktion simuliert. Eine KI-Anwendung, die automatisiert Vorlieben des*der Nutzer*in aus dem bisherigen Leseverhalten ermittelt und einen personalisierten Nachrichtenstrom aus Internet-basierten Inhalten erzeugt.
Gering	Die KI-Anwendung hat nur geringen Einfluss auf die Wahrnehmung oder Handlungen von Nutzer*innen oder Betroffenen. **Beispiele:** Eine KI-Anwendung, die Vogelstimmen erkennt oder Pflanzen bestimmt. Eine KI-Anwendung zur personalisierten Routenplanung. Eine KI-Anwendung zur personalisierten Planung touristischer Aktivitäten.

[AK-S] Dokumentation der Schutzbedarfsanalyse

Anforderung: Do

- Der Schutzbedarf der KI-Anwendung wird als *gering, mittel* oder *hoch* festgelegt. Die Wahl der Kategorie *gering/mittel/hoch* wird unter Bezugnahme auf die oben angeführte Tabelle ausführlich begründet.

Falls der Schutzbedarf für die Dimension Autonomie und Kontrolle *gering* ist, so ist keine nähere Betrachtung der einzelnen Risikogebiete erforderlich. Wurde hingegen ein *mittlerer* oder *hoher* Schutzbedarf ermittelt, so muss im Folgenden jedes Risikogebiet genauer untersucht werden.

5.1 Risikogebiet: Angemessene und verantwortungsvolle Gestaltung der Aufgabenverteilung zwischen Mensch und KI-Anwendung (GE)

KI-Anwendungen werden meist in Kontexten eingesetzt, in denen verschiedene Akteure (Firmen, Behörden, Verbraucherorganisationen, Betriebsräte, Datenschutzbeauftragte etc.) agieren. Dabei gehen sie oftmals komplexe Problemstellungen an, die Unsicherheit aufweisen und einen gewissen Grad an Systemautonomie erfordern. Das Bestreben, alle Anforderungen zu berücksichtigen und dabei den Menschen in den Mittelpunkt zu stellen, führt zu der politischen Diskussion, wie KI-Anwendungen angemessen und verantwortungsvoll eingesetzt werden sollen.

Der Fokus dieses Risikogebiets liegt auf der Gestaltung der Aufgabenverteilung zwischen Mensch und KI-Anwendung. Es soll sichergestellt werden, dass die KI-Anwendung aufgrund ihrer Aufgaben und ihres Autonomiegrads nicht ungerechtfertigt oder mit unvertretbarem Risiko die Handlungs- oder Wahrnehmungsfähigkeit von Nutzer*innen einschränkt. Stattdessen sollte der Betrieb der KI-Anwendung den Zweck haben, Menschen zu unterstützen und sie bestenfalls zu höheren oder anspruchsvolleren Tätigkeiten zu befähigen.

Die Autonomie der KI-Anwendung, d. h. wie »selbständig« bzw. (un-)überwacht das System agiert, steht im Spannungsfeld zur Autonomie des*der Nutzer*in. KI-Anwendungen mit hohem Autonomiegrad, die beispielsweise Entscheidungen treffen und darauf basierend, ohne Bestätigung durch Nutzer*innen, weiterführende Prozesse oder Aktionen einleiten, ersparen den beteiligten Personen in der Regel Zeit und Aufwand. Es sollte jedoch nicht außer Acht gelassen werden, dass ein hoher Autonomiegrad der KI-Anwendung zugleich den Handlungsspielraum von Nutzer*innen einschränkt. Daher ist individuell abhängig von Anwendungsbereich, Einsatzkontext und Verlässlichkeit der KI-Anwendung zu entscheiden, welcher Autonomiegrad für die jeweilige KI-Anwendung vertretbar und verantwortungsvoll ist.

An die Aufgabenverteilung zwischen Mensch und KI-Anwendung knüpfen weitere Fragestellungen an. Beispielsweise ist klarzustellen, welche Vorkenntnisse oder ggf. welches Expertenwissen für Menschen erforderlich ist, um eine korrekte Nutzung sowie eine effektive Überwachung bzw. Kontrolle der KI-Anwendung gewährleisten zu können. Außerdem sollte untersucht werden, inwiefern das Risiko übermäßigen Vertrauens in die KI-Anwendung (Stichwort *Automation Bias*) besteht und wie involvierte Personen angemessen darüber aufgeklärt werden können. Diese Themen werden im **Risikogebiet: Sicherstellung der Informiertheit und Befähigung von Nutzer*innen und Betroffenen (IB)** adressiert und auch in der **Dimension: Sicherheit (SI)**, unter anderem in Hinblick auf Unfallsituationen, aufgegriffen.

Insbesondere bei KI-Anwendungen, die die Sicherheit von Menschen gefährden können, beispielsweise die KI-basierte Steuerung eines Fahrzeugs, sind angemessene Eingriffsmöglichkeiten wie etwa die Kontrollübergabe an die Nutzer*innen vorzusehen. Die Einschränkung der Systemautonomie bei Verlassen des Normalzustands etwa im Sinne von Fehlertoleranz oder *Fail-Safe* wird im **Risikogebiet: Funktionale Sicherheit (FS)** adressiert (siehe auch **[SI-R-FS-MA-12]** zur Möglichkeit des menschlichen Eingriffs) und sollte mit den Zielvorgaben in diesem Risikogebiet konsistent sein.

5.1.1 Risikoanalyse und Zielvorgaben

[AK-R-GE-RI-01] Gestaltung einer Aufgabenverteilung zwischen Mensch und KI-Anwendung

Anforderung: Do

- **Risikoanalyse:** Es wird analysiert, inwiefern die KI-Anwendung (aufgrund ihres Zwecks aber auch aufgrund der Gestaltung der Aufgabenverteilung) Einfluss auf die Handlungs- und Wahrnehmungsfähigkeit von Nutzer*innen nehmen kann.
  - Zum einen wird angesichts Art und Umfang der möglichen, durch die KI-Anwendung zu übernehmenden Aufgaben untersucht, welche Handlungsoptionen oder -fähigkeiten dadurch bei Nutzer*innen oder anderen relevanten Personengruppen/Organisationen potenziell eingeschränkt werden. Beispielsweise könnte ein Spurhalteassistent, der orangefarbene Straßenmarkierungen nicht erkennt, den*die Fahrer*in daran

hindern, an einer Baustelle entlang der umgeleiteten Straßenführung zu fahren. Ein weiteres Beispiel ist, dass eine KI-Anwendung, deren Funktionalität etwa die Überwachung von Nutzerverhalten miteinschließt, dadurch die Handlung von Nutzer*innen beeinflusst.
  - Um die durch den Einsatz der KI-Anwendung entstehenden Abhängigkeiten zu ermitteln, werden die Konsequenzen einer teilweisen oder völligen Abschaltung der KI-Anwendung nach erfolgreicher Inbetriebnahme beschrieben. Dabei ist zu untersuchen, inwiefern die Notwendigkeit besteht, dass Nutzer*innen die Aufgaben der KI-Anwendung bei Ausfall kurzfristig übernehmen können und inwiefern Nutzer*innen der KI-Anwendung (auch langfristig gesehen) dazu in der Lage sind.

  Für jede identifizierte potenzielle Einschränkung der Handlungs- oder Wahrnehmungsfähigkeit involvierter Personengruppen wird abschließend abgeschätzt, welche (materiellen oder immateriellen) Schäden dies bei den Betroffenen zur Folge haben kann.
- **Zielvorgaben:** Es wird die angestrebte Aufgabenverteilung zwischen der KI-Anwendung und ihren Nutzer*innen beschrieben und unter Bezugnahme auf die Risikoanalyse begründet. Dabei sollten die Autonomie der KI-Anwendung und die Nutzerautonomie ausführlich gegeneinander abgewogen werden. Außerdem sind Zielvorgaben bezüglich der Überwachung und Kontrolle der KI-Anwendung zu formulieren, die erforderlich sind, um das Risiko der ungerechtfertigten Einschränkung von Nutzerautonomie bzw. Autonomie anderer involvierter Personengruppen zu senken. Falls zutreffend, ist dabei auch Bezug zu nehmen auf die Integration der KI-Anwendung in bestehende Arbeitsprozesse.

5.1.2 Kriterien zur Zielerreichung

[AK-R-GE-KR-01] Autonomiegrad der KI-Anwendung und Nutzerautonomie

Anforderung: Do

- Der Autonomiegrad einer KI-Anwendung kann grob in die folgenden vier Stufen[41] eingeteilt werden:
  - *Human Control* (HC)
    - Es handelt sich um ein reines Assistenzsystem. Die KI-Anwendung kann nicht ohne Bestätigung durch Nutzer*innen Anschlussaktionen auslösen.
    - Der Mensch trifft basierend auf der Ausgabe der KI-Anwendung eine Entscheidung bzw. leitet nächste Schritte ein; er ist in alle Entscheidungen involviert.

    Beispiele: *Forecasts* oder *Decision Support*-Systeme, die umfassende Information und Auswahl aufbereiten.
  - *Human-in-the-Loop* (HIL)
    - Die KI-Anwendung agiert teilautonom, sie kann aber keine Aufgabe ohne die Bedienung/Bestätigung durch einen Menschen abschließen.
    - Der Mensch hat umfassenden Über- und Einblick in die Operationen der KI-Anwendung, kann jederzeit in die relevanten Abläufe eingreifen und ist in die meisten Entscheidungsprozesse involviert. Insbesondere kann er Entscheidungen, die von der KI-Anwendung automatisch getroffen werden, nachträglich korrigieren, überschreiben und kompensieren.

    Beispiele: Personalisierte Empfehlungen, konkrete Vorschläge wie etwa Gesichtserkennung auf Fotos, Text- oder Emojivorschläge im Chat
  - *Human-on-the-Loop* (HOL)
    - Unter Normalbedingungen ist die KI-Anwendung (beinahe) in der Lage, autonom zu agieren bzw. Aufgaben ohne menschlichen Eingriff zu erledigen.
    - Der Mensch ist unter Normalbedingungen in keine oder nur wenige Entscheidungen involviert, hauptsächlich überwacht er die KI-Anwendung. Einschreiten ist nicht jederzeit und nicht an jeder Stelle möglich, aber der Mensch kann Entscheidungen, die von der KI-Anwendung automatisch getroffen werden,

---

41 Die Darstellung der Autonomiegrade ist angelehnt an: Nothwang, W., et al. (2016). The Human Should be Part of the Control Loop? In 2016 Resilience Week (RWS), pp. 214-220, IEEE https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7573336 (letzter Aufruf: 22.06.2021). Anmerkung: Der höchste Autonomiegrad, der im Paper »complete autonomy« genannt wird, ist in diesem Katalog als »human-out-of-the-loop« bezeichnet.

nachträglich korrigieren, überschreiben und kompensieren. Bei unerwarteten Ereignissen oder Fehlern ist menschlicher Eingriff erforderlich.

Beispiele: *E-Mail Spam Detector*, Entscheidung über automatisierte Kreditvergabe, *Fraud Detector* (sofern Menschen im Nachgang Entscheidungen der Anwendung überschreiben können; ansonsten würde es zu *Human-out-of-the-Loop* zählen)

– *Human-out-of-the-Loop* (HOOTL)
  · Die KI-Anwendung agiert unter allen Bedingungen (auch bei Fehlern oder unerwarteten Ereignissen) komplett autonom, d. h. es kann Aufgaben ohne menschlichen Eingriff vollständig erledigen; »*Sense-Think-Act*«.
  · Der Mensch kann nur noch entscheiden, ob er die KI-Anwendung einsetzen möchte oder nicht und u. U. das Setup/Metabefehle definieren (z. B. die Zieladresse bei einem autonomen Fahrzeug).

Beispiele: *High Frequency Trading*, Staubsaugerroboter

Der für die KI-Anwendung angestrebte Autonomiegrad ist gemäß den Zielvorgaben einer der beschriebenen Kategorien zuzuordnen. Des Weiteren ist zu spezifizieren, was der gewählte Autonomiegrad bezogen auf den vorliegenden Anwendungskontext konkret bedeutet und weshalb der Autonomiegrad angemessen ist.

- Mit Blick auf die Gestaltung der Aufgabenverteilung zwischen Mensch und KI-Anwendung (Eingriffsmöglichkeiten des Menschen, Involvierung in Entscheidungen) werden Anforderungen an die Nutzerautonomie festgehalten. Dabei sind die folgenden Punkte zu adressieren und durch anwendungsspezifische Vorgaben zu ergänzen:
  – Art und Umfang menschlicher Aufsicht/Überwachung der KI-Anwendung im Betrieb
  – Handlungsfreiheit von Nutzer*innen
  – Integration in den Arbeitsprozess, insbesondere mögliche Anforderungen bzgl. Kontrolle der Ausgaben durch Menschen
  – Beschwerdemöglichkeiten für Nutzer*innen und Betroffene
- Es ist zu begründen, dass die Kriterien mit den Zielvorgaben konform sind und untereinander keine Widersprüche aufweisen. Ferner ist darzulegen, dass bei Erfüllung der Kriterien die in **[AK-R-GE-RI-01]** identifizierten Risiken auf ein vertretbares Maß gesenkt werden.

5.1.3 Maßnahmen

Die folgenden Maßnahmen stehen losgelöst von Daten, KI-Komponente, Einbettung und Betrieb.

5.1.3.1 Daten

5.1.3.2 KI-Komponente

5.1.3.3 Einbettung

5.1.3.4 Maßnahmen für den Betrieb

[AK-R-GE-MA-01] Einbindung relevanter Personengruppen/Organisationen

Anforderung: Do

- Es liegt eine Dokumentation darüber vor, welche relevanten Personengruppen und Organisationen bei der Entwicklung der KI-Anwendung, insbesondere in Hinblick auf die Gestaltung der Aufgabenverteilung zwischen Mensch und KI-Anwendung beteiligt wurden. Dabei wird insbesondere beschrieben, ob und inwiefern
  – die in **[AK-R-GE-RI-01]** als relevant identifizierten Personengruppen/Organisationen (bzw. deren Vertreter*innen) eine Beschreibung der KI-Anwendung erhalten haben.

  - falls möglich, mehrere alternative Gestaltungsmöglichkeiten der KI-Anwendung ausgearbeitet und ihre Vor- und Nachteile gegeneinander abgewogen wurden.
  - Stellungnahmen der involvierten Personengruppen bzw. Organisationen eingeholt und die darin vorgebrachten Argumente abgewogen wurden.
- Es ist darzulegen, inwiefern die dokumentierten Maßnahmen zur Erfüllung der Kriterien **[AK-R-GE-RI-01]** beitragen.

[AK-R-GE-MA-02] Vorrang menschlichen Handelns

Anforderung: Do

- Es ist dokumentiert, in welchen Fällen ein*e Nutzer*in entscheiden kann, die KI-Anwendung nicht zu benutzen bzw. eine von der KI-Anwendung getroffene Entscheidung (direkt oder über bestimmte Ansprechpartner*innen) nachträglich zu korrigieren und ggf. zu kompensieren.
- Es liegt eine Dokumentation darüber vor, in welche Aktionen/Entscheidungen der KI-Anwendung Nutzer*innen während des (Normal-)Betriebs eingreifen können.
- Darüber hinaus wird dokumentiert, welche Eingriffsmöglichkeiten Nutzer*innen bei Abweichung vom Normalbetrieb zur Verfügung stehen oder unter Umständen gar erforderlich sind. Hierbei kann u. a. auf die entsprechenden Stellen im **Risikogebiet: Funktionale Sicherheit (FS)** (siehe auch **[SI-R-FS-MA-12]**) verwiesen werden.
- Falls Nutzer*innen zur Wahrnehmung der beschriebenen Eingriffsmöglichkeiten über bestimmte Qualifikationen/Kenntnisse verfügen müssen, so sind diese festzuhalten.
- Abschließend ist unter Bezugnahme auf die obigen Punkte zu erläutern, inwiefern dadurch ein Beitrag zur Realisierung des in **[AK-R-GE-KR-01]** vorgesehenen Autonomiegrads der KI-Anwendung geleistet wird. Außerdem ist in Bezug auf die obigen Punkte darzulegen, inwiefern die in **[AK-R-GE-KR-01]** definierten Anforderungen an die Nutzerautonomie erfüllt werden.

[AK-R-GE-MA-03] Etablierung wirksamer Beschwerdemöglichkeiten

Anforderungen: Do | Pr

- Es liegt eine Dokumentation vor, die belegt, dass Nutzer*innen und Betroffene Beschwerden über Einschränkungen ihrer Wahrnehmungs- und Handlungsfähigkeit durch die KI-Anwendung einbringen können. Es wird festgehalten, welche Gremien/Stellen für das Auswerten der Beschwerden zuständig sind. Der Prozess, um Beschwerden auszuwerten und daraus gegebenenfalls Konsequenzen abzuleiten, wird erläutert. Insbesondere wird beschrieben, inwiefern im Rahmen dieses Prozesses eine Abschaltung oder Weiterentwicklung der KI-Anwendung durchsetzbar ist.

[AK-R-GE-MA-04] Rollen- und Rechtekonzept für die Nutzung der KI-Anwendung

Anforderung: Do

- Es liegt ein Konzept über die Rollen und Verantwortlichkeiten hinsichtlich der Nutzung der KI-Anwendung vor. Darin ist beschrieben, welche Tätigkeiten im Rahmen der Nutzung, wie beispielsweise das Überschreiben/Korrigieren von Ausgaben der KI-Anwendung, einer Autorisierung bedürfen. Außerdem ist dokumentiert, wie sichergestellt wird, dass nur autorisierte Personen die entsprechenden Tätigkeiten ausüben.

[AK-R-GE-MA-05] Menschliche Aufsicht über die KI-Anwendung

Anforderungen: Do | Pr

- Es ist ein Prozess zur Überwachung und Kontrolle der KI-Anwendung durch Menschen etabliert. Das Ausmaß der Kontrolle sowie die Abläufe und Tätigkeiten innerhalb des Prozesses werden ausführlich beschrieben. Hierbei ist insbesondere auf die Maßnahmen zur Anwendungsüberwachung und zur Überwachung äußerer Faktoren Bezug zu nehmen, die in den Risikogebieten mit Titel »Beherrschung der Dynamik« in den übrigen Dimensionen beschrieben werden. Gegebenenfalls kann auf diese direkt verwiesen werden.
- Es liegt eine Dokumentation vor, die die Methoden und Werkzeuge beschreibt, mit denen ein*e Nutzer*in die korrekte Funktion der KI-Anwendung überprüfen kann. Eine korrekte Funktion beinhaltet die Einhaltung der

Zielvorgaben und die Beherrschung der Risiken aller in diesem Katalog behandelten Dimensionen. Insbesondere wird festgehalten, auf welche Weise Nutzer*innen erkennen können, wenn Fehler oder Abweichungen vom Normalbetrieb vorliegen und wie sie sich in diesem Fall zu verhalten haben. Falls dies bereits an anderer Stelle behandelt wurde, kann stattdessen ein entsprechender Verweis eingefügt werden.

- Ferner wird beschrieben, welche Qualifikationen oder weiteren Kenntnisse erforderlich sind, damit Nutzer*innen in der Lage sind, die KI-Anwendung effektiv zu beaufsichtigen und zu erkennen, wann ggf. ein Eingreifen erforderlich ist.
- Abschließend ist unter Bezugnahme auf die obigen Punkte zu erläutern, inwiefern dadurch ein Beitrag zur Realisierung des in **[AK-R-GE-KR-01]** vorgesehenen Autonomiegrads der KI-Anwendung geleistet wird. Außerdem ist in Bezug auf die obigen Punkte darzulegen, inwiefern die Anforderungen an die Nutzerautonomie erfüllt werden.

[AK-R-GE-MA-06] Abschalt-Szenarien

Anforderung: Do

- Es werden Szenarien identifiziert, analysiert und bewertet, in denen der Betrieb der KI-Anwendung zur Erhaltung der Wahrnehmungs- und Handlungsfähigkeit von Nutzer*innen und Betroffenen ganz oder teilweise abgeschaltet werden muss. Dabei werden nicht nur Abschaltungen aufgrund potenzieller Personen- und Sachschäden, sondern auch aufgrund der Verletzung von Persönlichkeitsrechten oder der Autonomie von Nutzer*innen und Betroffenen behandelt. Dementsprechend sind an dieser Stelle je nach Einsatzkontext Szenarien zu analysieren, die über die in der **Dimension: Sicherheit (SI)** behandelten Unfälle/Sicherheitsfälle hinausgehen. Beispielsweise könnte das Szenario betrachtet werden, dass eine KI-Anwendung zu Diskriminierung geführt hat und sich dies nicht umgehend beheben lässt. Bei der Bewertung der Szenarien sind auch die Konsequenzen des Abschaltens für involvierte Personen, Arbeitsprozesse, Organisation und Gesellschaft sowie zusätzliche Zeit- und Kostenaufwände zu dokumentieren. Dies wird den potenziellen Schäden aufgrund von Nicht-Abschalten der KI-Anwendung gegenübergestellt.
- Es liegt eine Dokumentation über die Strategien für eine Abschaltung der KI-Anwendung vor, die basierend auf den identifizierten Szenarien erarbeitet wurden – sowohl für eine kurzfristige, mittelfristige und dauerhafte Abschaltung. Ebenso sind Szenarien für die Abschaltung von Teilfunktionalitäten der KI-Anwendung zu dokumentieren. Unter Umständen wurden Abschalt-Szenarien bereits im **Risikogebiet: Funktionale Sicherheit (FS)** (siehe **[SI-R-FS-MA-10]**) thematisiert, auf die verwiesen werden kann. In einem Abschalt-Szenario ist dokumentiert,
  - das Setting und die daraus hervorgehenden Entscheidungsgründe für die Abschaltung,
  - die Dringlichkeit der Abschaltung,
  - durch welche Personen bzw. Rollen und auf welche Weise die Abschaltung durchgeführt wird,
  - wie der jeweilige Ausfall kompensiert werden kann,
  - welche Folgen für den*die Einzelne*n bzw. für die betroffene Organisation zu erwarten sind.

[AK-R-GE-MA-07] Technische Bereitstellung von Abschaltmöglichkeiten

Anforderung: Do

- Es liegt eine Dokumentation über die technischen Möglichkeiten vor, um einzelne Teil-Funktionalitäten der KI-Anwendung ebenso wie die gesamte KI-Anwendung abzuschalten. Hierbei kann ggf. auf **[SI-R-FS-MA-10]** oder **[SI-R-FS-MA-12]** Bezug genommen werden.
- Es wird dargelegt, dass andere Systemkomponenten oder Geschäftsprozesse, die eine abschaltbare (Teil-) Funktionalität verwenden, überprüft wurden, und (technische) Maßnahmen, die negative Auswirkungen von Abschaltungen kompensieren, vorbereitet sind. Falls dort bereits behandelt, kann dafür auf **[SI-R-FS-MA-10]** verwiesen werden.

5.1.4 Gesamtbewertung

[AK-R-GE-BW] Gesamtbewertung

Anforderung: Do

- Unter Bezugnahme auf die ergriffenen Maßnahmen wird zusammenfassend begründet, dass die in **[AK-R-GE-KR-01]** festgelegten Anforderungen, insbesondere der angestrebte Autonomiegrad sowie die Anforderungen an die Nutzerautonomie, realisiert wurden.
- Sofern nicht alle in **[AK-R-GE-KR-01]** spezifizierten Anforderungen erfüllt werden, sind die Abweichungen zu dokumentieren. Dies gilt ebenfalls für nur teilerfüllte Anforderungen, bei denen etwa die Kriterien nicht, oder nicht immer, erreicht wurden.

5.2 Risikogebiet: Sicherstellung der Informiertheit und Befähigung von Nutzer*innen und Betroffenen (IB)

Dieses Risikogebiet soll sicherstellen, dass Nutzer*innen und Betroffene selbstbestimmt und informiert mit der KI-Anwendung umgehen. Die »Informiertheit« im Titel des Risikogebiets bezieht sich dabei nicht vorrangig auf technische Funktionalitäten. Stattdessen wird die Offenlegung der KI-Anwendung, die Aufklärung von Nutzer*innen und Betroffenen über Rechte und Risiken und die korrekte Bedienung der KI-Anwendung thematisiert. Die technische Dimension selbstbestimmter Nutzung, d. h., dass Nutzer*innen beispielsweise die Entstehung einzelner Ausgaben verstehen und diese inhaltlich einordnen können, wird hingegen im **Risikogebiet: Transparenz gegenüber Nutzer*innen und Betroffenen (NB)** in der Dimension Transparenz behandelt.

Um eine ordnungsgemäße Nutzung der KI-Anwendung zu erreichen, muss sichergestellt werden, dass potenzielle Nutzer*innen eine hinreichende Erklärung zu Anwendungsbereich und –zweck sowie eine hinreichende und verständliche Erklärung über den korrekten Gebrauch der KI-Anwendung erhalten. Dazu gehört auch, dass den Nutzer*innen sowohl Autonomiegrad als auch Anforderungen an menschliche Überwachung/Kontrolle der KI-Anwendung kenntlich gemacht werden. Beispielsweise müssen Nutzer*innen einer KI-basierten Anwendung zur Bearbeitung von Kundenanfragen vor Versand der automatisiert erstellten Antwort informiert werden, falls diese Anwendung zur reinen Assistenz entwickelt wurde und deren Ausgaben durch menschliches Personal gegengeprüft werden sollen. Nutzer*innen müssen außerdem über Eingriffsmöglichkeiten in die KI-Anwendung sowie das erforderliche Verhalten im Ausnahmezustand informiert werden (siehe auch **Risikogebiet: Funktionale Sicherheit (FS)** in der Dimension Sicherheit). Gerade bei KI-Anwendungen mit hohem Autonomiegrad, bei denen unter Umständen nur kurze Zeitfenster bestehen, in denen Ausgaben überschrieben werden können, ist es wichtig, dass die zuständigen Personen die erforderlichen Handlungen beherrschen. Dementsprechend sollte auch kommuniziert werden, welche Qualifikationen für die korrekte Nutzung der KI-Anwendung notwendig sind, und wie diese erreicht werden können.

Auch bei korrekter Bedienung der KI-Anwendung können Risiken bestehen, die Nutzer*innen und Betroffene oder etwa die weitere Verarbeitung von Ergebnissen der KI-Anwendung betreffen. KI-Anwendungen, die direkt mit Menschen interagieren, bergen beispielsweise das Risiko, dass Nutzer*innen oder Betroffene eine emotionale Bindung aufbauen und dadurch manipulierbar und von der Anwendung (emotional) abhängig werden. Durch sogenanntes *Social Engineering* können Chatbots beispielsweise Nutzer*innen dazu veranlassen, bereitwillig persönliche Daten preiszugeben oder einen Großteil ihrer Freizeit für fiktive Interaktion zu opfern. Für KI-Anwendungen, die nicht den Zweck der Manipulation oder Täuschung, sondern der Verbesserung des Wohlbefindens haben, ist es natürlich nicht per se ausgeschlossen, dass sie gezielt die Gefühle der Nutzer*innen ansprechen. Als Beispiel ist dafür die KI-basierte Babyrobbe Paro[42] zu nennen, die in der Pflege eingesetzt wird, um u. a. demenzkranken Menschen Gesellschaft zu leisten und deren Stresslevel zu reduzieren. Jedoch sollten bei KI-Anwendungen, die mit Menschen interagieren und dabei zwangsläufig (gezielt oder unbeabsichtigt) Emotionen ansprechen, positive und negative Auswirkungen auf Nutzer*innen und Betroffene sorgfältig abgewogen werden.

Zusätzlich dazu sollte überprüft werden, ob die Gestaltung des Nutzer-Interfaces im Einklang mit Zweck und Fähigkeiten der KI-Anwendung steht. Das Aussehen, z. B. falls dies besonders menschenähnlich ist, sowie die Art der Ergebnisaufbereitung könnten beeinflussen, wie Nutzer*innen die Qualität und Zuverlässigkeit der KI-Anwendung wahrnehmen und ob sie der KI-Anwendung weitere Fähigkeiten wie etwa Empathie zuschreiben. Daraus ergibt sich das Risiko übermäßigen Vertrauens in die KI-Anwendung. Die Tendenz von Menschen, automatisierten Entscheidungen eher unkritisch gegenüberzustehen und sich nicht den Aufwand zu machen, diese durch andere Quellen zu überprüfen, wird auch als *Automation Bias* bezeichnet. Falls Nutzer*innen die Fähigkeiten der KI-Anwendung überschätzen, könnte dies etwa zur Folge haben, dass diese nicht mehr ausreichend überwacht wird und somit in einem höheren Autonomiegrad operiert als vorgesehen. Dieses Risiko sollte, ebenso wie das Risiko emotionaler Bindung und Manipulation, durch Information und Aufklärung von Nutzer*innen und Betroffenen abgeschwächt werden.

---

42 Baisch, S. et. al. (2018). Emotionale Roboter im Pflegekontext: Empirische Analyse des bisherigen Einsatzes und der Wirkungen von Paro und Pleo. *Z Gerontol Geriat* 51, 16–24. https://doi.org/10.1007/s00391-017-1346-8 (letzter Aufruf: 22.06.2021)

5.2.1 Risikoanalyse und Zielvorgaben

[AK-R-IB-RI-01] Risikoabschätzung

Anforderung: Do

- **Risikoanalyse:** Es wird analysiert, welche Risiken für die vorliegende KI-Anwendung in Hinblick auf unzureichende Informiertheit oder Befähigung von Nutzer*innen und Betroffenen bestehen und welche potenziellen Schäden daraus resultieren können. Die Analyse ist wie folgt strukturiert:
  - Es wird untersucht, welche Risiken sich durch unzureichende Informiertheit bzgl. des korrekten Gebrauchs der KI-Anwendung ergeben. Dabei ist zu berücksichtigen, inwiefern bzw. auf welche Weise die KI-Anwendung durch Nutzer*innen falsch bedient oder ungenügend überwacht werden könnte. Es ist sowohl der Normalbetrieb als auch der Ausnahmezustand zu betrachten, wobei bei Letzterem unter Umständen auf die Ausführungen im **Risikogebiet: Funktionale Sicherheit (FS)** verwiesen werden kann. Anschließend werden potenzielle Schäden abgeschätzt, die entstehen können, wenn die KI-Anwendung durch Menschen nicht wie vorgesehen genutzt/bedient/überwacht wird.
  - Ferner wird beschrieben, auf welche Weise die KI-Anwendung mit Nutzer*innen und Betroffenen interagiert und wie das Nutzer-Interface aussieht. Darauf basierend wird abgeschätzt, inwiefern involvierte Personen ein unangemessen hohes Vertrauen oder eine emotionale Bindung zu der KI-Anwendung aufbauen können. Außerdem wird analysiert, welche (materiellen oder immateriellen) Schadensszenarien bei unangemessen hohem Vertrauen oder emotionaler Bindung involvierter Personen auftreten können. Dabei ist mindestens zu untersuchen, inwiefern und zu welchen Themen Nutzer*innen bzw. Betroffene durch die KI-Anwendung manipuliert werden können.
  - Darüber hinaus wird untersucht, welche Schäden entstehen können, wenn Nutzer*innen und Betroffene nur unzureichend über das hinter der KI-Anwendung stehende Geschäftsmodell bzw. mögliche Interessen hinsichtlich der KI-Anwendung informiert werden.
- **Zielvorgaben:** Basierend auf der Risikoanalyse werden Zielvorgaben bezüglich der Informiertheit und Befähigung von Nutzer*innen und Betroffenen formuliert.
  Es wird dokumentiert, welche Informationen die Nutzer*innen und Betroffenen über die KI-Anwendung erhalten sollen. Dabei sind sowohl Informationen zur ordnungsgemäßen Nutzung als auch zur Aufklärung über die damit verbundenen Risiken abzudecken. Eine konkrete Auflistung wird in **[AK-R-IB-KR-02]** vorgenommen. Außerdem werden qualitative Ziele bezüglich der Sichtbarkeit der Informationen festgelegt. Es ist darzulegen, dass die Zielvorgaben bezüglich der Informationen unter Berücksichtigung und Abwägung der Interessen aller Beteiligten entsprechend dem Verhältnismäßigkeitsgrundsatz (z. B. Wahrung von Geschäftsgeheimnissen, öffentliches Interesse etc.) erfolgt ist und dem Kontext der KI-Anwendung angemessen ist. Gegebenenfalls werden Anforderungen an die Qualifikation von Nutzer*innen festgehalten, die sich aus der Risikoanalyse als Ergänzung zu den in **[AK-R-GE-MA-02]** dokumentierten Anforderungen ergeben. Diese können durch weitere Zielvorgaben bezüglich der Befähigung von Nutzer*innen und Betroffenen ergänzt werden. Die Wahl der Zielvorgaben ist zu begründen.

5.2.2 Kriterien zur Zielerreichung

[AK-R-IB-KR-01] Qualifikation von Nutzer*innen

Anforderung: Do

- Basierend auf den Zielvorgaben in **[AK-R-IB-RI-01]**, den Dokumentationen in **[AK-R-GE-MA-02]** und **[AK-R-GE-MA-05]** sowie unter Berücksichtigung von **[SI-R-BD-KR-01]**, werden die Anforderungen an die Qualifikation von Nutzer*innen formalisiert.

[AK-R-IB-KR-02] Vollständigkeit der Informationen für Nutzer*innen und Betroffene

Anforderung: Do

- Aufbauend auf der Risikoanalyse werden Inhalte/Informationen aufgelistet, die Nutzer*innen und Betroffenen vermittelt werden müssen, um eine ordnungsgemäße und selbstbestimmte Nutzung sowie ein Bewusstsein über alle relevanten Risiken zu ermöglichen. Falls bestimmte Personengruppen innerhalb der Nutzer*innen und Betroffenen unterschiedliche Informationen erhalten sollen, sind für diese Personengruppen separate Auflistungen zu erstellen. Beispielsweise sollten Mitarbeiter*innen einer Bank, die eine KI-Anwendung zur Kreditprüfung unmittelbar beaufsichtigen, Hinweise zur Bedienung und Funktionalität der KI-Anwendung erhalten, die womöglich für Kund*innen nicht relevant sind. Anhand der Auflistung wird in der Gesamtbewertung beurteilt, ob die Information für Nutzer*innen und Betroffene vollständig ist. Bei der Auflistung sollten mindestens die folgenden Punkte berücksichtigt werden:
  - Nutzer*innen und Betroffene werden darüber informiert, dass die KI-Anwendung im Einsatz ist.
  - Ein Verständnis des Zwecks und der Funktionen der KI-Anwendung wird vermittelt.
    - Nutzer*innen und Betroffenen ist es möglich, den Anwendungsbereich und –zweck der KI-Anwendung zu verstehen.
    - Eine präzise und vollständige Beschreibung der KI-Anwendung liegt vor. (»Vollständig« ist abhängig vom Einsatzkontext zu konkretisieren. Hierbei kann beispielsweise auf den **KI-Steckbrief (ST)** verwiesen werden.)
    - Es wird ein Bezug zu Informationen aus anderen Dimensionen hergestellt (z. B. werden Risiken, relevante Zielwerte etc. mitgeteilt). Insbesondere werden Nutzer*innen und Betroffene darüber informiert, wie vollständig, verlässlich und fair die durch die KI-Anwendung erlangten Ausgaben sind. Eine KI-Anwendung kann beispielsweise zu jeder Ausgabe auch eine Konfidenz-Aussage liefern, die mit verständlichen Handlungsempfehlungen für die Person verbunden sind. Ebenso können Hinweise gegeben werden, wie sich die Ausgaben der KI-Anwendung überprüfen lassen, z. B. durch Hinzuziehen anderer, unabhängiger Informationsquellen.
  - Ein Verständnis über die Geschäftsmodelle der Betreiber*innen und die Einbettung der KI-Anwendung in Prozesse wird vermittelt.
    - Das der KI-Anwendung zugrundeliegende Geschäftsmodell und sein Zweck werden erläutert.
    - Die Aufgabenverteilung zwischen der KI-Anwendung und ihren Nutzer*innen ist beschrieben.
    - Die mit der KI-Anwendung zusammenhängenden Arbeitsprozesse und die hierbei involvierten Personen bzw. Rollen (darunter auch Ansprechpartner*innen für Fragen) werden dargestellt (wie etwa in **[AK-R-GE-RI-01]**).
    - Die Konsequenzen einer teilweisen oder völligen Abschaltung der KI-Anwendung sind beschrieben (wie in **[AK-R-GE-MA-06]**).
    - Nutzer*innen und Betroffene sind darüber informiert, auf welchem Wege eine (teilweise) Abschaltung der KI-Anwendung durchgesetzt werden kann. Sie haben Kenntnis über die diesbezüglichen Entscheidungsgremien und -prozesse (siehe dazu auch **[AK-R-GE-MA-06]**, **[AK-R-GE-MA-07]**, **[SI-R-BD-MA-01]** und **[SI-R-BD-MA-03]**).
  - Informationen zur korrekten Nutzung und Beaufsichtigung sowie Eingriffsmöglichkeiten werden vermittelt. Die Möglichkeit zur Umsetzung der Handlungen wird sichergestellt.
    - Nutzer*innen und Betroffene können eine informierte Einwilligung oder Ablehnung der Anwendung treffen.
    - Nutzer*innen und Betroffene werden über Alternativen zur Nutzung der KI-Anwendung informiert. Falls beispielsweise die Möglichkeit besteht, ein Anliegen an Sachbearbeiter*innen zu richten, anstatt dieses automatisiert durch die KI-Anwendung bearbeiten zu lassen, werden die Nutzer*innen oder Betroffenen auf diese Option hingewiesen.
    - Nutzer*innen werden über den korrekten Gebrauch der KI-Anwendung informiert, z. B. im Rahmen einer Schulung.
    - Es liegt ein Nutzerhandbuch vor, das die KI-Anwendung, deren korrekte Nutzung und Beaufsichtigung sowie Eingriffsmöglichkeiten beschreibt.
      - Korrektheit der Beschreibung
      - Regelmäßige Reviews (mindestens einmal pro Jahr) und Updates des Handbuchs
      - Versionierung

        - Nutzer*innen sind in der Lage, falls erforderlich, Ausgaben der KI-Anwendung zu überschreiben oder gar während des Betriebs in Aktionen der KI-Anwendung einzugreifen. Insbesondere sollten sie die in **[AK-R-GE-MA-02]** und **[SI-R-FS-MA-12]** beschriebenen Eingriffsmöglichkeiten wahrnehmen können.
        - Nutzer*innen sind darüber informiert, wie die KI-Anwendung effektiv und angemessen beaufsichtigt bzw. kontrolliert wird. Dabei sollten sie mindestens über die in **[AK-R-GE-MA-05]** beschriebenen Abläufe zur Überwachung und Kontrolle der KI-Anwendung informiert werden.
        - Nutzer*innen wissen, wie sie sich in Ausnahmesituationen zu verhalten haben. Hierbei sollte über **[SI-R-FS-MA-12]**, **[SI-R-FS-MA-01]**, **[SI-R-IV-MA-01]**, **[SI-R-BD-MA-03]** und informiert werden.
    - Nutzer*innen und Betroffene werden über Risiken aufgeklärt.
        - Nutzer*innen und Betroffene werden über anwendungsspezifische Risiken in Bezug auf die Nutzerautonomie informiert, insbesondere in Fällen, in denen die KI-Anwendung suggestiv oder gar manipulativ in die Entscheidung von Personen eingreifen kann. Falls für die KI-Anwendung relevant, wird über das Risiko emotionaler Bindung/Abhängigkeit sowie über das Risiko übermäßigen Vertrauens in die KI-Anwendung (sog. *Automation Bias*) aufgeklärt.
        - Nutzer*innen und Betroffene sind über ihre Rechte und Beschwerdemöglichkeiten aufgeklärt. Dabei sollten sie mindestens über die in **[AK-R-GE-MA-03]** beschriebenen Beschwerdemöglichkeiten informiert werden.
- Falls einer der oben genannten Inhalte nicht an Nutzer*innen oder Betroffene vermittelt werden soll, ist dies zu begründen. Ferner kann die Auflistung durch weitere (anwendungsspezifische) Inhalte/Informationen ergänzt werden.

[AK-R-IB-KR-03] Sichtbarkeit und Zugänglichkeit der Informationen für Nutzer*innen

Anforderung: Do

- Es werden Anforderungen an die Sichtbarkeit und Zugänglichkeit der gemäß **[AK-R-IB-KR-02]** zu vermittelnden Informationen festgehalten. Dabei sollten mindestens die folgenden Punkte adressiert werden:
    - Der Grad, zu dem die Kenntnisnahme der Information eingefordert wird: Dieser reicht von abrufbarer Information (z. B. Gebrauchsanweisung) über Nutzungshinweise, die vor Gebrauch der KI-Anwendung gelesen und bestätigt werden müssen bis hin zu expliziten Einweisungen und Schulungen der Nutzer*innen sowie durchgeführten Lernkontrollen. Die Anforderungen diesbezüglich sollten im Einklang mit **[SI-R-BD-MA-01]** stehen.
    - Die Aufbereitung der Information: Diese sollte verständlich und an die in **[AK-R-IB-KR-01]** geforderte Qualifikation der involvierten Personen angepasst sein.
- Es ist zu begründen, dass die Kriterien in Einklang mit den Zielvorgaben in **[AK-R-IB-RI-01]** stehen.

5.2.3 Maßnahmen

Die folgenden Maßnahmen stehen losgelöst von Daten, KI-Komponente, Einbettung und Betrieb.

5.2.3.1 Daten

5.2.3.2 KI-Komponente

5.2.3.3 Einbettung

5.2.3.4 Maßnahmen für den Betrieb

[AK-R-IB-MA-01] Aufbereitung der Informationen für Nutzer*innen und Betroffene

Anforderungen: Do | Pr

- Es wird ausführlich beschrieben, auf welche Weise die für Nutzer*innen und Betroffene relevante Information aufbereitet wird. Mögliche Arten der Aufbereitung relevanter Themen sind:
  - Ein Kommunikationsprozess, der die Nutzer*innen und Betroffenen darauf hinweist, dass sie mit einer KI-Anwendung kommunizieren bzw. die Entscheidungen auf einer KI-Anwendung beruhen. Auf einen solchen Hinweis kann verzichtet werden, wenn dies im Rahmen der Anwendung und unter Beachtung der Verhältnismäßigkeit begründet wird.
  - Publikation von Beschwerdemöglichkeiten, u. U. differenziert nach den Dimensionen der Vertrauenswürdigkeit, mit menschlichen Ansprechpartner*innen.
  - Eine für die Nutzer*innen verständliche Beschreibung für den korrekten Gebrauch der KI-Anwendung (Benutzerhandbuch). Aus dieser sollen auch der Einsatzbereich und -zweck der Anwendung, die vorgesehene Zielgruppe, die Aufgabenverteilung zwischen der KI-Anwendung und Nutzer*innen sowie die menschlichen Eingriffsmöglichkeiten klar hervorgehen.
  - Eine für Nutzer*innen und Betroffene einsehbare Dokumentation, die besondere Risiken und relevante Metriken aus anderen Dimensionen, insbesondere aus Sicherheit, Datenschutz, Verlässlichkeit und Fairness, aufzeigt und verständlich in den Kontext des Anwendungsgebietes setzt.
  - Sicherheitsrichtlinien (siehe **[SI-R-FS-MA-01]**, **[SI-R-IV-MA-01]**)
  - Notfallhandbuch (siehe **[SI-R-BD-MA-03]**)
- Es ist darzulegen, dass die beschriebenen Aufbereitungsmaßnahmen alle in **[AK-R-IB-KR-02]** gelisteten Inhalte/Informationen abdecken sind.
- Ferner wird beschrieben, auf welche Weise bzw. durch welche Kanäle die aufbereitete Information den Nutzer*innen und Betroffenen bereitgestellt wird. Falls erforderlich, wird zudem dargelegt, auf welche Weise sichergestellt wird, dass die Nutzer*innen und Betroffenen die gemäß **[AK-R-IB-KR-02]** für sie relevanten Informationen zur Kenntnis nehmen.
- Es ist zu begründen, wie diese Maßnahmen zur Erfüllung von **[AK-R-IB-KR-03]** beitragen.

[AK-R-IB-MA-02] Befähigung von Nutzer*innen

Anforderung: Do

- Es wird dokumentiert, auf welche Weise sichergestellt wird, dass die Nutzer*innen gemäß den Anforderungen **[AK-R-IB-KR-01]** qualifiziert sind. Dabei kann unter anderem auf **[SI-R-BD-MA-01]** Bezug genommen werden.

5.2.4 Gesamtbewertung

[AK-R-IB-BW] Gesamtbewertung

Anforderung: Do

- Es liegt eine Dokumentation vor, die zusammenfassend nachweist, dass die Kriterien **[AK-R-IB-KR-01]** bis **[AK-R-IB-KR-03]** durch die ergriffenen Maßnahmen erfüllt werden.
- Sofern nicht alle in **[AK-R-IB-KR-01]** bis **[AK-R-IB-KR-03]** spezifizierten Anforderungen erfüllt werden, sind die Abweichungen zu dokumentieren. Dies gilt ebenfalls für nur teilerfüllte Anforderungen, bei denen etwa die Kriterien nicht, oder nicht immer, erreicht wurden.

Zusammenfassende Betrachtung

[AK-Z] Zusammenfassende Betrachtung der Dimension

Anforderung: Do

- Falls für diese Dimension ein mittlerer oder hoher Schutzbedarf besteht, ist eine Dokumentation über die verbleibenden Restrisiken zu erstellen. Zunächst werden die Restrisiken aus den verschiedenen Risikogebieten dieser Dimension zusammengefasst. Anschließend wird unter Berücksichtigung des Schutzbedarfs analysiert, ob die identifizierten Restrisiken insgesamt als vernachlässigbar, nicht vernachlässigbar (aber vertretbar) oder unvertretbar zu bewerten sind. Das Ergebnis der Analyse ist zu erläutern.
- Falls potenziell negative Auswirkungen von Risiken oder Maßnahmen dieser Dimension auf andere Dimensionen festgestellt wurden, sind diese zu dokumentieren.
- Es wird ein Fazit über die Dimension gezogen, welches insbesondere die Bewertung der Restrisiken enthält.

6. Dimension: Transparenz (TR)

Beschreibung und Zielsetzung

Durch Künstliche Intelligenz können Problemstellungen angegangen werden, die mit klassischer Softwareentwicklung unlösbar schienen. Außerdem liefern maschinelle Lernverfahren in vielen Domänen bessere und präzisere Ergebnisse als herkömmliche Software, die auf fest einprogrammierten Regeln basiert. Neben diesen Vorteilen haben die meisten KI-basierten Anwendungen durch das datengetriebene Lernen jedoch einen Nachteil gegenüber der klassischen Softwareentwicklung: Das Zustandekommen ihrer Ausgaben ist selbst für KI- und Domänen-Experten*innen oftmals nur schwer nachvollziehbar. Die Eigenschaft einer KI-Anwendung, dass deren Funktionsweise und Entscheidungen vollständig oder teilweise von einem Menschen nachvollzogen werden können, wird als Transparenz oder Nachvollziehbarkeit bezeichnet. Dabei wird darin unterschieden, ob eine Erklärung für das Zustandekommen einer einzelnen Prädiktion geliefert wird oder ob das verwendete Maschinelle Lernverfahren als Ganzes transparent ist. Während Ersteres im Folgenden als Erklärbarkeit bezeichnet wird, bezeichnen wir Letzteres im Folgenden als Interpretierbarkeit des Modells. Beispielsweise gelten lineare Modelle im Allgemeinen als interpretierbar, da sie die Elemente der Eingabe lediglich mit einem Faktor gewichten, um eine Ausgabe zu erzeugen. Eine korrekte Normalisierung vorausgesetzt, lässt sich beispielsweise einfach ableiten, welche Teile der Eingabe den größten Einfluss auf die Ausgabe haben. Eine hinreichende (semantische) Interpretierbarkeit des linearen Modells im Sinne dieser Dimension erfordert zudem die Interpretierbarkeit der Features.

Die Dimension Transparenz bezieht sich auf die inneren Prozesse der KI-Anwendung und speziell auf das ML-Modell. Dabei werden Erklärbarkeit, Interpretierbarkeit, Nachverfolgbarkeit und Reproduzierbarkeit auf technischer Ebene untersucht. Nicht-technische Aspekte der Transparenz, z. B. ob sich die KI-Anwendung als solche zu erkennen gibt, werden in der **Dimension: Autonomie und Kontrolle (AK)** behandelt. Für die Dimension Transparenz ist ferner zu beachten, dass sie unter Umständen in Konflikt mit anderen Dimensionen wie zum Beispiel **Dimension: Sicherheit (SI)** oder **Dimension: Verlässlichkeit (VE)** stehen kann. Auf diese Problematik wird in der abschließenden, dimensionsübergreifenden Beurteilung der KI-Anwendung näher eingegangen.

In vielen Anwendungskontexten mag Transparenz eine untergeordnete Rolle spielen. Beispielsweise ist es kaum einem*einer Nutzer*in einer KI-basierten Spracherkennung wichtig zu wissen, wieso die KI-Anwendung das jeweilig gesprochene Wort richtig erkannt hat oder nicht. In solchen Anwendungskontexten, in denen die (technische) Transparenz der KI-Anwendung nicht sicherheitskritisch ist, dient Transparenz vorrangig zur Stärkung der Vertrauenswürdigkeit der KI-Anwendung oder sorgt für größere Zufriedenheit unter den Nutzer*innen. In anderen Bereichen ist die Erklärbarkeit von Ausgaben hingegen essenziell für die sichere und verantwortungsvolle Nutzung der KI-Anwendung. Zum Beispiel liefert eine KI-basierte Bilderkennung, die Ärzt*innen bei der Diagnose anhand eines MRT- oder Röntgenbilds unterstützen soll, nur dann einen echten Mehrwert für die Ärzt*innen, wenn diese die Ursachen der KI-basierten Entscheidung nachvollziehen können, indem beispielsweise als krankhaft erkanntes Gewebe im Bild markiert wird. In anderen Anwendungskontexten wie etwa einem KI-basierten Kreditscoring-System in einer Bank könnte die Interpretierbarkeit von Ausgaben die Einzelfallentscheidung von Sachbearbeiter*innen erleichtern, falls etwa potenzielle Kund*innen gegen eine schlechte Bewertung ihrer Kreditwürdigkeit Einspruch erheben.

Eine Möglichkeit, um Transparenz für Nutzer*innen zu erreichen, ist durch den Einsatz eines perse interpretierbaren Modells (im Gegensatz zu Black-Box-Modellen), welches grundsätzlich einen einfachen bzw. greifbaren Zusammenhang zwischen Eingangsgrößen und Ausgaben darstellt. Da interpretierbare Modelle jedoch nicht für alle Problemstellungen eine optimale Lösung darstellen, wird oftmals auf komplexere Black-Box-Modelle, wie etwa tiefe Neuronale Netze, zurückgegriffen. Zwar lassen sich die Berechnungen von Black-Box-Modellen rein algorithmisch ebenfalls nachvollziehen, die Bedeutung und Übertragbarkeit auf menschliche Begrifflichkeiten und Operationen

ist allerdings nicht gegeben. Für viele Black-Box-Modelle, die insbesondere nicht interpretierbar sind, kann jedoch zumindest Erklärbarkeit hergestellt werden, indem das Zustandekommen ihrer Ausgaben durch aufwendige, nachgeschaltete Verfahren wie z. B. das Trainieren von Surrogat-Modellen oder einer LIME-Analyse (*Local Interpretable Model-agnostic Explanations*) erklärt werden. Zurzeit ist die Erklärbarkeit[43] von Modellen ein aktives Forschungsfeld und es werden viele Anstrengungen unternommen, die Lernprozesse von Black-Box-Modellen besser zu verstehen sowie ihre internen Prozesse zu visualisieren und die resultierenden Entscheidungen zu erklären. Aus technischer Sicht ist die Frage der Transparenz nicht trivial und das Spannungsfeld zwischen Verlässlichkeit bzw. Robustheit der KI-Komponente und der Nachvollziehbarkeit ihrer Funktionsweise (etwa durch Wahl eines weniger performanten, aber dafür interpretierbaren Modells) ist ein altbekanntes Dilemma.[44]

Nicht nur für die Zufriedenheit von Nutzer*innen oder in Bezug auf die sichere oder verantwortungsvolle Nutzung einer KI-Anwendung spielt Transparenz eine Rolle. So sind Aspekte wie Nachverfolgbarkeit bzw. Reproduzierbarkeit von Ergebnissen auch für rechtliche Fragen relevant, z. B. bezüglich der Haftung im Fall unerwarteten Verhaltens der KI-Anwendung. Zwar müssen Expert*innen in sicherheitskritischen Anwendungskontexten nicht jede Ausgabe einer KI-Anwendung vorhersagen können, ihr generelles Verhalten muss jedoch während der Entwicklung und auch später im produktiven Betrieb erklärbar, nachvollziehbar und dokumentiert sein. Hierzu dienen Logdaten, Dokumentationen bzw. Archivierungen des Designs, der Daten, des Trainings, des Testens und Validierens des Modells, sowie der einbettenden Umgebung. Dies erlaubt es z. B. bei (interner) Revision, nach Feedback oder Beschwerden das Modell zu verbessern.

Aus diesen Vorüberlegungen ergeben sich folgende Risikogebiete:

1. **Transparenz gegenüber Nutzer*innen und Betroffenen:** Dieses Risikogebiet befasst sich mit Risiken, die dadurch entstehen, dass Entscheidungen und Auswirkungen der KI-Anwendung gegenüber Nutzer*innen und Betroffenen nicht hinreichend erklärt werden können.

2. **Transparenz für Expert*innen:** Dieses Risikogebiet widmet sich Risiken, die dadurch entstehen, dass das Verhalten der KI-Anwendung von einem*einer Expert*in nicht hinreichend verstanden und nachvollzogen werden kann.

3. **Auditfähigkeit:** Dieses Risikogebiet behandelt Risiken, die dadurch entstehen, dass die Entwicklung sowie die im Einzelfall ausgeführten Vorgänge im Betrieb der KI-Anwendung nicht hinreichend dokumentiert und belegt sind.

4. **Beherrschung der Dynamik:** Dieses Risikogebiet behandelt Risiken, die dadurch entstehen, dass sich Anforderungen an die Transparenz oder die implementierten Transparenzverfahren selbst ändern.

Schutzbedarfsanalyse

Je nach Anwendungskontext kann es eine essenzielle Voraussetzung für die sichere und verantwortungsvolle Nutzung einer KI-Anwendung sein, dass Nutzer*innen oder Expert*innen tieferen Einblick in die technischen Eigenschaften des Modells erlangen und das Zustandekommen seiner Ausgaben nachvollziehen bzw. interpretieren können. So fordert etwa auch der Verordnungsentwurf zur Künstlichen Intelligenz der Europäischen Kommission, dass in bestimmten Fällen das Zustandekommen einer Entscheidung für Nutzer*innen nachvollziehbar sein muss, um eine angemessene Nutzung zu ermöglichen.

---

43 Eine vertiefende technische Betrachtung hier diskutierter Bereiche und Methoden findet sich beispielsweise in: Samek, W., Montavon, G., Vedaldi, A., Hansen, L., Müller, K. (2019). Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Springer https://link.springer.com/book/10.1007/978-3-030-28954-6 (letzter Aufruf: 21.06.2021)

44 Die Darstellung in diesem Abschnitt ist stark angelehnt an das Kapitel »3.3 Transparenz« des Whitepapers: Poretschkin, M., et al. (2019). Vertrauenswürdiger Einsatz von Künstlicher Intelligenz. Sankt Augustin: Fraunhofer-Institut für Intelligente Analyse und Informationssysteme IAIS. https://www.ki.nrw/wp-content/uploads/2020/03/Whitepaper_KI-Zertifizierung.pdf (letzter Aufruf: 18.06.2021)

Das potenzielle Schadensszenario für diese Dimension der Vertrauenswürdigkeit ist dadurch charakterisiert, dass aufgrund von Intransparenz keine sichere und zweckmäßige Nutzung der KI-Anwendung möglich ist oder die KI-Anwendung gegen maßgebliche Vorgaben verstößt. Hieran orientiert sich auch die Bemessung des Schutzbedarfs für diese Dimension. Die Schadenskategorien richten sich danach, inwiefern die verschiedenen Aspekte der Transparenz zum einen für die ordnungsgemäße, d. h. sichere und verantwortungsvolle Nutzung sowie zum anderen für die Zweckmäßigkeit der KI-Anwendung, insbesondere, falls diese nicht sicherheitskritisch ist, relevant sind. Dazu wird unter anderem untersucht, ob Intransparenz zeitliche oder finanzielle Aufwände verursacht (z. B. um Erklärungen für Ausgaben zu finden, um eine Intransparenz der KI-Anwendung zu kompensieren), oder ob die KI-Anwendung bei Intransparenz ohne Beeinträchtigungen genutzt werden kann.

Beispiel: Ein Online-Fashion-Store setzt eine KI-Anwendung als Recommender-System ein, welches je Kund*in ein persönliches Recommender-Profil erstellt. Eine Information über die Funktionsweise der KI-Anwendung und die verarbeiteten Daten ist für Kund*innen transparent einzusehen, weshalb diese das Zustandekommen der Empfehlungen selbst gegenprüfen können. Dem*der Betreiber*in des Online-Shops entstehen somit weniger finanzielle Aufwände, da es weniger Rücksendungen gibt. Außerdem ist der Service für Kund*innen des Online-Shops durch die Nachvollziehbarkeit des Recommender-Systems besonders ansprechend und zufriedenstellend, wodurch sich für den*die Betreiber*in ein Wettbewerbsvorteil und Gewinn von Kund*innen ergibt.

Die Tatsache, dass sich potenziell Trade-Offs zu anderen Dimensionen wie etwa **Dimension: Sicherheit (SI)** oder **Dimension: Verlässlichkeit (VE)** ergeben können, wird in der dimensionsübergreifenden Beurteilung der KI-Anwendung adressiert und sollte nicht in die Ermittlung des Schutzbedarfs einfließen.

Der Schutzbedarf wird folgendermaßen kategorisiert:

Hoch	Es gibt eine Transparenzanforderung (Erklärbarkeit, Interpretierbarkeit oder Nachverfolgbarkeit/Reproduzierbarkeit), bei deren Nichterfüllung die KI-Anwendung ▪ entweder für den ursprünglich zugedachten Zweck unbrauchbar wäre, z. B., weil eine sichere oder verantwortungsvolle Nutzung nicht möglich erscheint, ▪ oder nur unter (zeitlich oder finanziell unvertretbarem) zusätzlichen Aufwand zweckmäßig betrieben werden könnte. Ein hohes Schadenspotenzial besteht außerdem bereits dann, wenn durch Intransparenz gegen maßgebliche (gesetzliche/normative) Vorgaben verstoßen würde. **Beispiel:** Eine KI-Anwendung, die eine medizinische Diagnose stellt, aber nicht nachvollziehbar ist.
Mittel	Es können Situationen auftreten, in denen das Nichterfüllen einer Transparenzanforderung den Nutzen der KI-Anwendung mindert und es (zeitlichen oder finanziellen) Aufwands bedarf, um die Zweckmäßigkeit/den Nutzen der KI-Anwendung herzustellen. Gleichzeitig ist es nicht möglich, dass durch Intransparenz gegen maßgebliche (gesetzliche/normative) Vorgaben verstoßen würde. **Beispiel:** Eine KI-Anwendung, die von einem Unternehmen in Kundenangelegenheiten (z. B. Fragen nach dem Kreditrahmen) eingesetzt wird. Falls es eine Anfrage zu den Gründen für eine bestimmte Ausgabe gibt, so wird deren aus Kundensicht zufriedenstellende Beantwortung durch eine Intransparenz des Modells erschwert. **Beispiel:** Eine KI-Anwendung, die Bilder in sozialen Netzwerken automatisiert nach unangemessenem Inhalt beurteilt und ggf. deren Veröffentlichung blockiert. Falls Bilder blockiert werden, ohne dass die für die Ausgabe ausschlaggebenden Bildbereiche markiert sind oder der vermeintliche Verstoß kategorisiert wird, erhöht dies den Zeitaufwand für menschliche Operatoren, falsch positiv oder falsch negativ eingeordnete Bilder zu identifizieren und nachzuprüfen.

Gering	Es gibt keinen Aspekt der Transparenz, dessen Nichterfüllung die Sicherheit oder die Zweckmäßigkeit der KI-Anwendung mindern könnte; oder aber es sind lediglich geringe Auswirkungen auf die Zweckmäßigkeit der KI-Anwendung möglich, die durch wiederum geringen Aufwand behoben werden können. Gleichzeitig ist es nicht möglich, dass durch Intransparenz gegen maßgebliche (gesetzliche/normative) Vorgaben verstoßen würde. **Beispiel:** Eine KI-basierte Zugangskontrolle, die die Zutrittsberechtigung von Personen über eine kamerabasierte Gesichtserkennung ermittelt. Im Fall einer Fehlklassifikation wäre eine Angabe der Gesichtsmerkmale, die zur Entscheidung geführt haben, nicht sinnvoll, da sie einem*einer Expert*in keinen Mehrwert bei der Korrektur der Entscheidung gibt. Vielmehr wird ein menschlicher Kontrolleur sein visuelles System und sein implizites oder explizites Berufswissen über Personenidentifikation nutzen, um die Entscheidung der KI-Anwendung zu überprüfen.

[TR-S] Dokumentation der Schutzbedarfsanalyse

Anforderung: Do

- Der Schutzbedarf der KI-Anwendung für die Dimension Transparenz wird als *gering, mittel* oder *hoch* festgelegt. Die Wahl der Kategorie *gering/mittel/hoch* wird unter Bezugnahme auf die oben angeführte Tabelle ausführlich begründet.

Falls der Schutzbedarf für die Dimension Transparenz *gering* ist, so ist keine nähere Betrachtung der einzelnen Risikogebiete erforderlich. Wurde hingegen ein *mittlerer* oder *hoher* Schutzbedarf ermittelt, so muss im Folgenden jedes Risikogebiet genauer untersucht werden.

6.1 Risikogebiet: Transparenz gegenüber Nutzer*innen und Betroffenen (NB)

Durch dieses Risikogebiet soll sichergestellt werden, dass den Nutzer*innen eine angemessene und verständliche Erklärung/Interpretation der Funktionsweise der KI-Anwendung zur Verfügung steht, um diese sicher, ordnungsgemäß und verantwortungsvoll bedienen oder verwenden zu können.

Transparenz kann beispielsweise durch die Wahl eines interpretierbaren Modells oder die verständliche Visualisierung von Erklärungen der Ergebnisse erreicht werden. Hierbei gibt es verschiedene Arten von Transparenzverfahren, die u. a. deutlich machen, welche der Input-Faktoren besonders ausschlaggebend für die jeweilige Ausgabe der KI-Anwendung sind. So kann zum Beispiel Ärzt*innen als Nutzer*innen einer KI-Anwendung zur Krankheitsdiagnose durch eine sogenannte Heatmap die Sensitivität der verschiedenen Input-Parameter im Blutbild angezeigt werden. Während einem medizinischen Laien und insbesondere den Patient*innen als Betroffenen diese Spezifikation unverständlich bliebe, kann sie für die behandelnden Ärzt*innen einen sinnvollen Beitrag zur Diagnose stellen. Transparenz sollte daher im Kontext der adressierten Zielgruppe betrachtet werden, beziehungsweise für unterschiedliche Zielgruppen passgenaue Herangehensweisen aufweisen. Der technische Detailgrad der Erklärung ist insbesondere an die anzunehmende Qualifikation der Nutzer*innen oder der Betroffenen anzupassen.

Neben dem spezifischen Einsatzkontext der KI-Anwendung können sich auch aus den internen Zielen des*der Betreiber*in weitere Transparenzanforderungen ergeben. Beispielsweise können sich Unternehmen/Organisationen als eigenes Leitbild eine starke Transparenz der eigenen Verfahren gegenüber Kund*innen setzen. Durch erweiterte Transparenzmaßnahmen, die über das notwendige Maß hinausgehen, können solche Unternehmen/Organisationen das Vertrauen der Kund*innen in ihre KI-Anwendungen stärken (Stichwort »Vertrauen durch Transparenz«).

6.1.1 Risikoanalyse und Zielvorgaben

[TR-R-NB-RI-01] Risikoanalyse und Grad der Transparenz der KI-Anwendung

Anforderung: Do

- **Risikoanalyse:** Es ist zu analysieren, welche potenziellen Schäden oder Gefährdungen entstehen können, wenn die KI-Anwendung intransparent gegenüber Nutzer*innen oder Betroffenen ist. Dabei sind verschiedene Grade von Transparenz zu betrachten. Außerdem ist die Eintrittswahrscheinlichkeit und -häufigkeit der identifizierten potenziellen Schäden abzuschätzen.
- **Zielvorgaben:** Basierend auf der Risikoanalyse sowie unter Bezug auf den Grundsatz der Verhältnismäßigkeit und unter Beachtung der typischen Qualifikation der Nutzer*innen und Betroffenen in diesem Anwendungsgebiet werden qualitative Zielvorgaben für die Transparenz der KI-Anwendung gegenüber diesen formuliert. Insbesondere wird beschrieben, für welchen Teil der KI-Anwendung welcher Grad an Transparenz, im Sinne von Interpretierbarkeit und Erklärbarkeit, als angemessen erachtet und angestrebt wird und weshalb.

6.1.2 Kriterien zur Zielerreichung

[TR-R-NB-KR-01] Bewertung der Erklärbarkeit gegenüber Nutzer*innen und Betroffenen

Anforderung: Do

- Es werden qualitative oder quantitative Kriterien definiert und erläutert, anhand derer der Grad der Transparenz der KI-Anwendung gegenüber Nutzer*innen und Betroffenen sinnvoll beurteilt werden kann. Falls für die gegenüber Nutzer*innen und Betroffenen zu erklärenden Sachverhalte bzw. Zusammenhänge (z. B. die Daten, das Modell, das Zustandekommen der Ergebnisse) unterschiedliche Kriterien angewendet werden sollen, so ist die Wahl des Kriteriums pro Sachverhalt bzw. Zusammenhang einzeln zu begründen. Die Wahl der Kriterien kann sich an der folgenden Auflistung qualitativer Bewertungsmaßstäbe orientieren. Falls

andere, qualitative oder quantitative, Kriterien festgelegt werden, sind diese zu beschreiben und deren Auswahl zu begründen.
  - Die Eindeutigkeit und Verständlichkeit der Erklärungen, die Nutzer*innen und/oder Betroffenen bezüglich der zu erklärenden Sachverhalte (z. B. Funktionsweise der KI-Anwendung, Daten, etc.) zur Verfügung gestellt wird, wurde durch einen signifikanten Anteil von herangezogenen Testpersonen bestätigt.
  - Die Eindeutigkeit und Verständlichkeit der Erklärungen, die Nutzer*innen und/oder Betroffenen bezüglich des Zustandekommens einzelner Ausgaben bzw. Ergebnisse der KI-Anwendung zur Verfügung gestellt wird, wurde durch einen signifikanten Anteil von herangezogenen Testpersonen bestätigt.
  - Die Qualifikation der herangezogenen Testpersonen entspricht der Qualifikation der erwarteten Zielgruppe bzw. den erwarteten Nutzer*innen und/oder Betroffenen.
  - Zur Bewertung der Erklärung einer Ausgabe, die Nutzer*innen und/oder Betroffenen gegebenen wird (z. B. durch ein Transparenzverfahren o. ä. erzeugt), können die folgenden Kriterien herangezogen werden:
    - Die Erklärung ist korrekt, in dem Sinne, dass das Transparenzverfahren bzw. die Erklärung der KI-Anwendung treu sind (z. B. sollten sich, falls vorhanden, Fehler des Modells in der Erklärung entsprechend widerspiegeln)
    - Die Erklärung trifft eine Aussage darüber, welche Input-Faktoren für die jeweilige Ausgabe ausschlaggebend/besonders wichtig sind
    - Die Erklärung ist bezüglich einer signifikanten Anzahl ähnlich gelagerter Fälle stabil/übertragbar/konsistent
    - Die Erklärung bereitet eine Konfidenz-Ausgabe des ML-Modells für Nutzer*innen und/oder Betroffene verständlich auf (vgl. **Risikogebiet: Einschätzung von Unsicherheit (UN)** in der **Dimension: Verlässlichkeit (VE)**)
    - Die Erklärung ermöglicht es dem*der Nutzer*in bzw. dem*der Betroffenen, »neue« Datenpunkte (sog. *Out-of-Domain*-Daten, im Sinne von Datenpunkten, die nicht in der vorgesehenen Verteilung liegen) als solche zu erkennen

    Beispiel: Ein ML-Modell wurde mithilfe von Nahaufnahmen der menschlichen Augen trainiert, Erkrankungen wie z. B. grauer Star zu erkennen. Ein Transparenzverfahren macht es möglich, die für die Vorhersage wichtigen Bildbereiche hervorzuheben. »Neue« Datenpunkte im Sinne von *Out-of-Domain*-Daten könnten hierbei Aufnahmen eines anderen Körperteils des Menschen sein, die somit nicht in der vorgesehenen Verteilung der Eingabedaten liegen. Hingegen sind Augen-Aufnahmen von Patient*innen, die nicht Teil der Trainingsdaten waren, zwar ebenfalls »unbekannt« für das ML-Modell, aber keine *Out-of-Domain*-Daten. In zahlreichen Anwendungsfällen, die etwa Rohdaten eines Sensors verarbeiten, kann eine derartige Unterscheidung zwischen »unbekannten« Daten für menschliche Nutzer*innen nur schwer zu treffen sein.
- Es wird ausführlich begründet, dass die hier festgelegten Kriterien die in **[TR-R-NB-RI-01]** formulierten Zielvorgaben abbilden.

6.1.3 Maßnahmen

6.1.3.1 Daten

[TR-R-NB-MA-01] Trainings- und Testdaten

Anforderung: Do

- Es liegt eine Dokumentation der Trainings- und Testdaten vor, aus der ersichtlich ist, um welche Art von Daten es sich handelt und ob diese für vorgesehene Nutzer*innen oder Betroffene vor dem Hintergrund etwaiger Vorkenntnisse verständlich/interpretierbar sind (bezogen auf die intrinsische Interpretierbarkeit der Daten). Falls die Daten nicht intrinsisch interpretierbar sind, so muss eine Begründung erfolgen, weshalb gerade diese Art von Daten verwendet werden. Ferner sind die Maßnahmen zu beschreiben, die ergriffen wurden, um Nutzer*innen und Betroffenen ein grundlegendes Verständnis der Daten zu ermöglichen, beispielsweise durch Dokumentation oder Anleitung über die Daten.
  Beispiel: Bilddaten, in denen das von der KI-Anwendung zu untersuchende Phänomen gut erkennbar ist, sollten als verständlich angesehen werden, während die Rohdaten eines Sensors, gerade wenn sie

hochdimensional sind, nicht ohne großen Aufwand auf bestimmte Zustände zurückgeführt werden können, also im Rahmen dieser Maßnahme als nicht verständlich angesehen werden.
- Sofern die Daten vor Verwendung als Trainings- bzw. Testdaten einer Vorverarbeitung unterzogen werden, wie beispielsweise einer Filterung, Bereinigung oder Transformation, ist zu dokumentieren. Ferner ist zu prüfen, ob weitere Erläuterungen dieser Vorverarbeitungsschritte für die Verständlichkeit der KI-Anwendung für Nutzer*innen und/oder Betroffene erforderlich sind. Vorgesehene Erklärungen oder aber eine Begründung bezüglich ihres Verzichts sind dokumentiert.

6.1.3.2 KI-Komponente

[TR-R-NB-MA-02] Interpretierbarkeit des ML-Modells

Anforderung: Do

- Es liegt eine Dokumentation darüber vor, welche Modelle zur Bearbeitung der vorliegenden Problemstellung in Betracht gezogen wurden. Außerdem ist darzulegen, dass diese bezüglich ihrer Interpretierbarkeit bewertet wurden, z. B. durch nähere Untersuchung der Wahl der Architektur, also der möglichen Operationen und deren möglicher Reihenfolge.
- Es liegt eine Dokumentation vor, in der beschrieben wird, weshalb man sich für das gewählte Modell (bzw. die Architektur) entschieden hat. Falls kein interpretierbares Modell verwendet wird, muss ersichtlich sein, weshalb kein solches Modell eingesetzt wird. Eine Auswahl von zum Zeitpunkt der Veröffentlichung des Katalogs als interpretierbar geltenden Modelle ist nachfolgend gelistet.

Zu den interpretierbaren Modellen[45] zählen (Stand Juli 2021):
- Lineare und logistische Regression (sowie deren Erweiterungen wie *Generalized Linear Models* und *Generalized Additive Models*)
- Lineare und logistische Klassifikation
- Entscheidungsbäume
- Regellernen
- naiver Bayes'scher Klassifikator
- k-nächste Nachbarn

[TR-R-NB-MA-03] Nachvollziehbarkeit der Funktionsweise

Anforderung: Do

- Es liegt eine Dokumentation und/oder eine Visualisierung des Modells (ggf. einschließlich schematischer Abbildung der Architektur) mit angemessenen Erklärungen vor, um die Funktionsweise nachvollziehbar zu machen. Es ist zu beschreiben, wie diese Visualisierung bzw. Erklärung den Nutzer*innen und Betroffenen der KI-Anwendung zugänglich gemacht wird.

[TR-R-NB-MA-04] Zustandekommen der Ergebnisse

Anforderung: Do

- Es liegt eine Dokumentation vor, in der die ergriffenen Maßnahmen beschrieben werden, mit denen Erklärungen über das Zustandekommen von Ergebnissen erzeugt werden. Art und Umfang der erzeugten Erklärungen sind immer im Sinne der Verhältnismäßigkeit in Bezug auf die Qualifikation von Nutzer*innen und Betroffenen sowie in Bezug auf deren Nutzung der KI-Anwendung zu wählen. Beispielsweise könnten Erklärungen für Nutzer*innen sinnvoll sein, für Betroffene derselben KI-Anwendung aber unverhältnismäßig.

---

45 Gemäß: Molnar, C. (Juni 2021). Chapter 4 Interpretable Models | Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. https://christophm.github.io/interpretable-ml-book/simple.html (letzter Aufruf: 16.06.2021)

- Im Fall intrinsisch-interpretierbarer Modelle:
  - Für regelbasierte Modelle ist anzugeben, welche Regeln für die aktuelle Entscheidung greifen (z. B. könnte bei Entscheidungsbäumen die jeweils gewählte Verästelung neben dem Ergebnis mit angegeben werden).
  - Bei Modellen der linearen Regression und verwandten Modellen sind die für eine Entscheidung relevantesten bzw. ausschlaggebendsten Attribute anzugeben (bei normalisierten Eingaben der größte Absolutwert der Koeffizienten).
- Bei Black-Box-Modellen:
  Es sind Methoden zu beschreiben, die zur Generierung von Erklärungen oder zur Förderung der Interpretierbarkeit ergriffen wurden. Die Methoden sollten entweder modellspezifisch sein, oder aber Modell-agnostisch anwendbar, wie etwa die nachfolgend gelisteten Methodenbeispiele:
  - *Partial Dependence Plots*
  - *Individual Conditional Expectation Plots*
  - *Accumulated Local Effects (ALE) Plot*
  - *Feature Interaction*
  - *Feature Importance*
  - *Global Surrogate Model*
  - *Student-Teacher-Model* mit interpretierbarem *Student*
  - *Local Surrogate (LIME)*
  - *Shapley Values*

  Zulässig sind auch die folgenden Beispiel-basierten Erklärungen (meist Modell-agnostisch, aber auf einzelne Datenpunkte bezogen statt auf den kompletten Eingabe-Output-Zusammenhang):
  - *Counterfactual Explanations*
  - *Prototypes*
  - *Influential Instances*

[TR-R-NB-MA-05] Statistische Evaluation der Erklärungen

Anforderungen: Do | Te

- Es werden Tests durchgeführt und dokumentiert, die nachweisen, dass die Erklärungen der Ausgaben der KI-Anwendung, wie sie in **[TR-R-NB-MA-04]** dargestellt werden, die in **[TR-R-NB-KR-01]** geforderten Eigenschaften erfüllen. Hierbei sollte die Treue der Erklärungen berücksichtigt werden, insbesondere im Fall von Fehlermodi. Die Art des Tests sollte sich hierbei stark an der gewählten Erklärungsmethode orientieren. Beispielsweise könnte zur Erklärung einer Bildklassifikation ein Heatmap-Verfahren herangezogen werden. In diesem Fall ließe sich zum einen prüfen, ob die Hervorhebung der Heatmap auf für die Klassifikation (nach menschlichem Ermessen) relevante Bildteile entfällt und inwieweit die Heatmap bei Fehlklassifikation noch auf spezifische Bereiche lokalisiert ist.
- Teile dieses Tests können mit Maßnahmen zur Überprüfung von Anforderungen aus dem **Risikogebiet: Transparenz für Expert*innen (EX)** übereinstimmen, vgl. etwa **[TR-R-EX-KR-02]**. Sofern Tests identisch sind oder sich überschneiden kann an dieser Stelle auf die entsprechenden Maßnahmen referenziert werden.

6.1.3.3 Einbettung

[TR-R-NB-MA-06] Kommunikation der Begründungen von Entscheidungen

Anforderung: Do

- Es liegt eine Dokumentation vor, aus der ersichtlich wird, in welchen Fällen bzw. nach welchen Kriterien die KI-Anwendung die Gründe ihrer Entscheidungen nach außen kommuniziert. Außerdem ist die Art der Kommunikation zu beschreiben (z. B. Anzeigen oder Visualisieren der Erklärungen aus **[TR-R-NB-MA-04]**). Wenn keine Kommunikation der Entscheidungen stattfindet oder die Kommunikation nicht in allen Fällen stattfindet, so muss in Bezug auf den Verhältnismäßigkeitsgrundsatz und die durchgeführte Risikoanalyse begründet werden, weshalb dies nicht passiert bzw. weshalb dies nicht möglich ist.

[TR-R-NB-MA-07] Menschliche Evaluation der Erklärungen

Anforderungen: Do | Te

- Es wird getestet und dokumentiert, dass die Anforderungen an die Eindeutigkeit und Verständlichkeit von Erklärungen für Nutzer*innen und Betroffenen, wie in **[TR-R-NB-KR-01]** definiert, sowohl bezüglich Darstellung als auch Inhalt erfüllt werden. Hierzu kann eine Studie mit menschlichen Tester*innen durchgeführt und evaluiert werden. Art und Fragestellungen der Studie sind im Detail zu beschreiben und deren Wahl zu begründen. Sowohl die Durchführung der Studie, inklusive der Wahl der Teilnehmer*innen, ihrer Anzahl und Qualifikation, als auch die Ergebnisse der Studie werden dokumentiert. Das Qualifikationsniveau der Studienteilnehmer sollte in etwa demjenigen der zu erwartenden und/oder angestrebten Zielgruppe entsprechen. Dies kann durch Referenzieren auf die entsprechenden Teilabschnitte der Studie erfolgen.

6.1.3.4 Maßnahmen für den Betrieb

[TR-R-NB-MA-08] Prozess zur Beantwortung von Benutzeranfragen

Anforderungen: Pr | Te

- Es wird ein zuvor empirisch getesteter Prozess etabliert, der gewährleistet, dass Nutzer*innen und Betroffene der KI-Anwendung nach Anfrage Erklärungen für die Ausgaben der KI-Anwendung erhalten können. Dieser Prozess ist zu dokumentieren. Alternativ muss eine Begründung gegeben werden, warum ein solcher Prozess für die Zielerreichung nicht benötigt wird. Hierzu kann ggf. auch auf die Informationen, die den Nutzer*innen und Betroffenen gemäß **[TR-R-NB-MA-03]** und **[TR-R-NB-MA-06]** standardmäßig bereitgestellt werden, verwiesen werden.

6.1.4 Gesamtbewertung

[TR-R-NB-BW] Gesamtbewertung

Anforderung: Do

- Es liegt eine Dokumentation vor, in der erläutert wird, dass aufgrund der getroffenen Maßnahmen die in **[TR-R-NB-KR-01]** definierten Kriterien erfüllt werden.
- Sofern nicht alle in **[TR-R-NB-KR-01]** spezifizierten Anforderungen erfüllt werden, sind die Abweichungen zu dokumentieren. Dies gilt ebenfalls für nur teilerfüllte Anforderungen, bei denen etwa die Kriterien nicht, oder nicht immer, erreicht wurden.

6.2 Risikogebiet: Transparenz für Expert*innen (EX)

Das Risikogebiet Transparenz für Expert*innen ist eng verwandt mit dem Ziel im **Risikogebiet: Transparenz gegenüber Nutzer*innen und Betroffenen (NB)**. Allerdings liegt der Fokus im vorliegenden Risikogebiet auf der Validierung, etwa zum Aufdecken von Modellschwächen, sowie auf der (technischen) Nachvollziehbarkeit und Reproduzierbarkeit von Ausgaben der KI-Anwendung durch Expert*innen. Das technische Level ist entsprechend höher.

In vielen Fällen, etwa bei der Verwendung Neuronaler Netze, ist es selbst für KI-Expert*innen schwer nachzuvollziehen, wie Ausgaben zu Stande kommen. In Bereichen mit hohen Anforderungen an Zuverlässigkeit und/oder Nachvollziehbarkeit stellt dies ein potenzielles Risiko dar. Ziel dieses Risikogebiets ist es daher, mit introspektiven Methoden

- Entscheidungen nachvollziehbar zu machen,
- Ausgaben zu plausibilisieren,
- die inhärente »Logik« der KI-Anwendung zu prüfen,
- und potenzielle Fehlerursachen sowie ggf. systematische Modellschwächen aufzudecken.

Die Erfüllung dieser Ziele kann dabei nicht nur zur Transparenz, sondern auch zur **Dimension: Verlässlichkeit (VE)** der KI-Anwendung beitragen.

Für Verfahren und Ansätze mit geringer Komplexität, z. B. *Decision Trees* oder *Clustering*, ist Transparenz und Interpretierbarkeit meist intrinsisch gegeben. Im Rahmen einer komplexen KI-Anwendung sind jedoch alle vier genannten Ziele mit enormem Aufwand verbunden und es ist möglich, dass in solchen Fällen aufgrund von Trade-Offs zwischen Aufwand und Umfang der Introspektionsmaßnahmen einerseits und Notwendigkeit und Zweckmäßigkeit andererseits eine vollständige Transparenz nicht erfüllt werden kann. Daher ist in der nachfolgenden Risikoanalyse zu klären, welcher Grad an Transparenz für Expert*innen bezüglich der KI-Anwendung angestrebt wird.

6.2.1 Risikoanalyse und Zielvorgaben

[TR-R-EX-RI-01] Risikoanalyse und Zielvorgaben

Anforderung: Do

- **Risikoanalyse:** Es wird analysiert, welche möglichen Schäden und Gefährdungen aufgrund mangelnder Transparenz der KI-Anwendung für Expert*innen sowie insbesondere aufgrund mangelnder Validierbarkeit bzw. Plausibilisierbarkeit der Ausgaben der KI-Anwendung durch Expert*innen auftreten können. Dabei werden sowohl die Eintrittswahrscheinlichkeit bzw. -häufigkeit, als auch die potenzielle Schadenshöhe abgeschätzt.
- **Zielvorgaben:** Basierend auf der Risikoanalyse werden qualitative Zielvorgaben bezüglich der verschiedenen Aspekte der Transparenz einer KI-Anwendung (Interpretierbarkeit, Erklärbarkeit und Nachvollziehbarkeit) gegenüber Expert*innen formuliert.
  Beispiel 1: Die Fußgängererkennung eines selbstfahrenden Fahrzeugs sollte hohen Ansprüchen an Zuverlässigkeit genügen. Im Schadensfall kann es erforderlich sein, die Entscheidungen der KI-Anwendung in hohem Detailgrad nachzuvollziehen.
  Beispiel 2: Eine KI-Anwendung zur Hochwasservorhersage modelliert physikalische Vorgänge. Zur Prüfung ihrer Zuverlässigkeit ist es daher sinnvoll, die Ausgaben bezüglich einfacher (physikalischer) Zusammenhänge zu plausibilisieren und zu validieren.

6.2.2 Kriterien zur Zielerreichung

Das komplexe Anforderungsbild im Bereich Transparenz stellt eine Herausforderung dar, wenn es darum geht, die Erfüllung der Zielvorgaben zu überprüfen. Für eine exakte Überprüfbarkeit sind hierbei, wo möglich, quantitative Vorgaben qualitativen vorzuziehen. Sofern mehr als eine Anforderung existiert, ist hierbei auch ein Wechselspiel zwischen quantitativen und qualitativen Vorgaben möglich. Maßgeblich ist jedoch, dass die Kriterien für alle Anforderungen separat und gleichzeitig erreichbar sein müssen. Ausnahmen hiervon, wie etwa Kriterien, die die gleiche Transparenzanforderung abbilden, sind gesondert zu begründen.

[TR-R-EX-KR-01] Anforderungen an die Eigenschaften von Transparenz-/Introspektionsverfahren

Anforderung: Do

- Es liegt eine Dokumentation vor, in der die Kriterien beschrieben werden, die zur Bewertung der Verfahren zur Herstellung von Transparenz/Introspektion hinzuzuziehen sind. Dabei sollte jedes Kriterium zum einen spezifizieren, auf welchen Sachverhalt bzw. Zusammenhang, der gegenüber Expert*innen erklärbar/interpretierbar/nachvollziehbar sein soll, sich dieses bezieht. Zum anderen sollte es Anforderungen an das Transparenzverfahren enthalten, durch das der entsprechende Sachverhalt bzw. Zusammenhang gegenüber Expert*innen erklärt wird. Dabei sind insbesondere zu erreichende Zielintervalle (quantitativ) bzw. zu erreichende Zieleigenschaften (qualitativ, strukturell) vorzugeben. Bei der Wahl der Kriterien sollten mindestens die folgenden Punkte aufgegriffen und erörtert werden:
  - Umfang, Ausgestaltung und Detailgrad der Transparenzverfahren. Insbesondere ist zu untersuchen, ob der im Anwendungskontext notwendige Grad an Transparenz durch das Verfahren erreicht wird.
  - Tiefe und Breite der Introspektion in Bezug auf die Modellausgaben. Hierbei ist zu erörtern, ob für jede Modellausgabe (Breite) ein Introspektionsverfahren anzuwenden ist, und für welche Modelltiefe die Introspektionsverfahren anzuwenden sind (z. B. Transparenz nur für finale Ausgaben des Modells oder auch für Zwischenergebnisse innerhalb des Modells).
  - zeitlicher Rahmen, innerhalb dessen eine Erklärung/Introspektion zur Verfügung stehen muss. Hier ist insbesondere zu beachten, ob Echtzeitanforderungen an die Introspektionsverfahren bestehen.
  - Komplexität des Transparenzverfahrens, d. h. wie aufwendig die Umsetzung des Verfahrens sein darf.
- Es wird ausführlich begründet, dass die festgelegten Kriterien separat und gleichzeitig erfüllt werden können, es also keinen Zielkonflikt gibt. Anforderungen, die nur gemeinsam aussagekräftig sind, um das Erreichen eines Ziels zu beurteilen, sollten hierbei zu einem gemeinsamen Kriterium zusammengefasst werden.
- Für jedes Kriterium und dessen zugehörige Zielwerte bzw. qualitative Zieleigenschaften wird begründet, dass diese dem Anwendungskontext angemessen und mit den Zielvorgaben in **[TR-R-EX-RI-01]** konform sind.

[TR-R-EX-KR-02] Anforderungen an die Ausgaben bzw. Ergebnisse von Transparenz-/Introspektionsverfahren

Anforderung: Do

- Die in **[TR-R-EX-KR-01]** festgelegten Anforderungen an die Eigenschaften von Transparenz- und/oder Introspektionsverfahren werden in Bezug auf deren Ergebnisse erweitert. Die Wahl der Kriterien ist zu dokumentieren und begründen. Hierbei sollten die folgenden Aspekte in Betracht gezogen werden:
  - die Stabilität der Erklärung bezüglich ähnlich gelagerter Fälle
  - die Korrelation zwischen der Konfidenz des Modells und dem Zutreffen der Erklärung
  - Verständlichkeit der Ergebnisse des Transparenzverfahrens für Expert*innen (bestätigt durch Testpersonen)
  - eine intrinsische Gewichtung der Aussage, d. h. ob eine oder mehrere Erklärungen mit einer Gewichtung in Form einer Konfidenz versehen sind
  - Treue, d. h. die möglichst geringe Abweichung der durch die Erklärung suggerierten Prädiktion von der tatsächlichen Prädiktion

  Diese Liste ist nicht vollständig. Zur Definition der Kriterien sollten auch für den vorliegenden Anwendungskontext spezifische Anforderungen einbezogen werden.

  Beispiel: Mögliche Kriterien zur Beurteilung der Ausgaben eines Heatmap-Verfahrens, das von Expert*innen verwendet werden soll, um die Ausgaben einer Bildklassifikation zu plausibilisieren, sind:
  - Das Heatmap-Verfahren soll auf ähnlichen Input-Bildern eine ähnliche Erklärung bieten (Stabilität).

  - Die Heatmap stellt den Einfluss bestimmter Bildregionen dar und kann daher relevante Bildfeatures hervorheben. Diese können von Menschen in Bezug auf die Klassifikation interpretiert werden.
  - Wird das Eingabebild verfremdet, sollten sowohl die Konfidenz des ML-Modells als auch die Stärke der Heatmap abnehmen, vgl. auch **Risikogebiet: Einschätzung von Unsicherheit (UN)** in der **Dimension: Verlässlichkeit (VE)**.
- Es wird ausführlich begründet, dass die genannten Kriterien separat und gleichzeitig erfüllt werden können, es also keinen Zielkonflikt gibt. Anforderungen, die nur gemeinsam zum Erreichen eines Ziels herangezogen werden können, sollten hierbei zu einem gemeinsamen Kriterium zusammengefasst werden.
- Es wird belegt, dass bei der Auswahl der Transparenzkriterien in ausreichendem Umfang Domänenwissen einbezogen wurde. Dies kann durch die Einbindung von Domänenexpert*innen auf dem Zielgebiet der KI-Anwendung erfolgen. Sofern kein solches Wissen für die Auswahl der Transparenzkriterien erforderlich ist, ist dies zu begründen.
  Beispiel: Auf Rat medizinischen Fachpersonals entscheidet man sich, eine KI-Anwendung zur Analyse eines Blutbildes in ihren Ausgaben zu plausibilisieren, indem die Relevanz der jeweiligen Eingabeparameter bestimmt wird.
- Für jedes Kriterium und dessen zugehörige Zielwerte bzw. qualitative Zieleigenschaften wird begründet, dass diese dem Anwendungskontext angemessen und mit den Zielvorgaben in **[TR-R-EX-RI-01]** konform sind. Bei qualitativen Kriterien ist zusätzlich zu erklären, warum kein quantitatives Kriterium gewählt wurde.

6.2.3 Maßnahmen

Die Zielsetzung bezüglich introspektiver Maßnahmen unterscheidet sich je nach zugrundeliegender KI-Anwendung und ihrem Einsatzkontext. Aus diesem Grund werden mögliche Herangehensweisen bestenfalls frühzeitig in der Entwicklung der KI-Anwendung als richtungsweisende Elemente in Betracht gezogen, vgl. **[TR-R-EX-MA-02]**. Bei der Transparenz von Maschinellen Lernverfahren handelt es sich um ein aktives Forschungsfeld, dessen aktueller Stand an dieser Stelle nur als Momentaufnahme zum Zeitpunkt der Veröffentlichung des Prüfkatalogs wiedergegeben werden kann. Eine tiefergehende Auseinandersetzung mit dem zum Zeitpunkt der Prüfung geltenden State of the Art ist insbesondere für Anwendungen mit hohem Transparenzbedarf unvermeidlich.

6.2.3.1 Daten

[TR-R-EX-MA-01] Eignung der Trainings- und Testdaten

Anforderung: Do

- Es liegt eine Dokumentation der Trainings- und Testdaten des ML-Modells vor, aus der ersichtlich ist, um welche Art von Daten es sich handelt und ob diese für vorgesehene Expert*innen verständlich/interpretierbar sind (bezogen auf die intrinsische Interpretierbarkeit der Daten). Falls die Daten nicht intrinsisch interpretierbar sind, so muss eine Begründung erfolgen, weshalb gerade diese Art von Daten verwendet werden. Ferner sind die Maßnahmen zu beschreiben, die ergriffen wurden, um Expert*innen das erforderliche Verständnis der Daten zu ermöglichen, beispielsweise durch Dokumentation oder eine Anleitung über die Daten. Ggf. kann hier auf **[TR-R-NB-MA-01]** verwiesen werden.
- Die ggf. zum Training sowie zum Testen der introspektiven Methoden verwendeten Daten werden dokumentiert und deren Wahl bzw. Eignung begründet. Je nach verwendetem Ansatz können die Anforderungen höher sein, als es für das reine Training der KI-Komponente erforderlich wäre.
  Zum Beispiel können *Out-of-Domain*-Daten hoher Qualität benötigt werden, um eine hinreichende Aussagekraft der durchgeführten Tests zu begründen. Außerdem könnte ein detaillierteres Labeling für Korrelationsanalysen bezüglich des Einsatzzwecks erforderlich sein, oder nachvollziehbare Datenquellen, um die zugrundeliegende Domäne besser einzugrenzen. Hierbei sind Überschneidungen mit den Anforderungen aus der **Dimension: Verlässlichkeit (VE)** möglich, und falls dort bereits behandelt, kann für die Wahl der Testdaten auf die entsprechende Stelle in der Dimension Verlässlichkeit verwiesen werden.

– Weiterhin kann es notwendig sein, Metadaten zu den Testdaten bereitzustellen, welche die Interpretierbarkeit der genutzten Test- und Trainingsdaten und/oder auch des verwendeten Transparenzverfahrens erhöhen. Dies ist insbesondere bei Transparenzverfahren nach Art eines visuell interaktiven Interfaces (vgl. **[TR-R-EX-MA-06]**) der Fall.

6.2.3.2 KI-Komponente

Gegeben dem Fall, dass durch die in **[TR-R-EX-KR-01]** und **[TR-R-EX-KR-02]** definierten Kriterien die Transparenz/Introspektion bezüglich unterschiedlicher Sachverhalte bzw. Zusammenhänge gefragt ist, so sind die Maßnahmen **[TR-R-EX-MA-03]** bis **[TR-R-EX-MA-05]** für jeden dieser Sachverhalte bzw. Zusammenhänge separat anzuwenden.

[TR-R-EX-MA-02] Begründete Wahl von Introspektions-/Transparenzmethoden

Anforderung: Do

- Es liegt eine Dokumentation vor, die eine Auseinandersetzung mit dem aktuellen Stand der Technik zur Realisierung von nachvollziehbarem Maschinellen Lernen in Bezug auf die KI-Anwendung und den vorliegenden Einsatzkontext nachweist. Im Allgemeinen kann Transparenz auf verschiedenen Ebenen einer KI-Anwendung hergestellt werden, z. B. durch Analyse der Ausgaben aber auch durch die Definition interner Zustände oder Hilfsgrößen wie etwa Gradienten. Es wird dargelegt, welche technischen Möglichkeiten zum Erreichen von Transparenz der KI-Anwendung und zur Plausibilisierung ihrer Ausgaben betrachtet wurden. Der Fokus der Darstellung ist nach den in **[TR-R-EX-RI-01]** definierten Zielvorgaben auszurichten.
  Mögliche Ansätze zur Introspektion, die für die Umsetzung der Transparenzanforderungen in Betracht gezogen werden können, sind folgende:
  – Analyse von Grenzflächen, etwa bei Klassifikation
  – Heatmaps auf der Ausgabe (bspw. DeconvNet)
  – Analyse latenter Darstellungen und/oder Repräsentationen (bspw. TCAV)
  – Zusammenhang zwischen Korrelation oder Kausalität (vgl. *Brittle Features*)
  – Verwendung interpretierbarer lokaler Proxymodelle (bspw. LIME)
  – Sensitivitätsanalysen bezüglich Parameter, sowohl des Modells als auch des Inputs
  – Betrachtung der verwendeten Verlustfunktion
  – Auswertung von »*Attention*«
  Diese Verfahren unterscheiden sich sowohl nach Zielsetzung als auch zugrundeliegender ML-Modelle (z. B. Random Forests, SVMs oder DNNs). Es besteht kein Anspruch auf Vollständigkeit.
- Ferner ist zu dokumentieren, inwiefern sich die betrachteten Transparenzmethoden auf die Verlässlichkeit und Leistungsfähigkeit der KI-Anwendung auswirken. Für jede in Betracht gezogene Herangehensweise/Methode zur Herstellung von Transparenz werden diejenigen Eigenschaften dargelegt, bei denen mit einer derartigen Beeinflussung zu rechnen ist.
  Beispiel: Eine Anforderung an die Interpretierbarkeit der latenten Darstellung eines Autoencoders, die in den Loss (bspw. durch einen Regularisierungsterm) eingreift, führt in der Regel zu größeren bis mittleren Rekonstruktionsfehlern.
- Es wird dokumentiert, welche Herangehensweisen bzw. Methoden zur Umsetzung der in **[TR-R-EX-RI-01]** formulierten Transparenzanforderungen zusätzlich zu den für Nutzer*innen und Betroffene zur Verfügung gestellten Erklärungsmethoden (siehe Maßnahme **[TR-R-NB-MA-04]**) in der KI-Anwendung implementiert sind.
  – Basierend auf der Auseinandersetzung mit dem aktuellen Stand der Technik und mit den Auswirkungen von Transparenzmethoden auf die Verlässlichkeit und Leistungsfähigkeit, wird die Wahl der implementierten Herangehensweisen bzw. Transparenzmethoden ausführlich begründet.
  – Falls zutreffend, wird dargelegt, wie bei der Entwicklung (durch die Berücksichtigung von Transparenzanforderungen im Design, oder auch durch die Implementierung von Methoden zur Herstellung von Transparenz) mit negativen Auswirkungen auf andere Dimensionen (insbesondere **Dimension: Verlässlichkeit (VE)**) umgegangen wurde.

  - Außerdem wird erläutert, inwiefern die gewählten Herangehensweisen bzw. Methoden (»by Design« oder als nachträglich implementierte Transparenzverfahren) die Kriterien in **[TR-R-EX-KR-01]** erfüllen.
- Es ist zu erörtern, ob eine aggregierte Betrachtung der gewählten Introspektionsmethoden, etwa mithilfe eines visuell-interaktiven Interfaces (siehe **[TR-R-EX-MA-06]**), sinnvoll und anwendbar ist. Werden die Introspektionsmethoden einzeln, d. h. ohne aggregierte Betrachtung, angewendet, ist diese Entscheidung zu begründen.

[TR-R-EX-MA-03] *Sanity Check* der implementierten Herangehensweise/Transparenzmethode

Anforderungen: Do | Te

- Es liegt eine Dokumentation eines Tests (sog. »*Sanity Check*«) vor, der die Effektivität bzw. Plausibilität der implementierten Herangehensweise belegt. Genauer gesagt soll durch den Test überprüft werden, ob die implementierte Herangehensweise tatsächlich Einblick in die spezifische Funktionsweise des ML-Modells ermöglicht, anstatt etwa eine Erklärung zu erzeugen, die de facto unabhängig vom Modell ist. Die Gestaltung des Tests[46] und die Wahl des Testdatensatzes ist mit Bezug auf die KI-Anwendung zu begründen. Die Durchführung sowie die Ergebnisse des Tests sind zu dokumentieren.
  Beispiel: Ein Heatmap-Verfahren soll eine KI-Anwendung zur Klassifikation von Bildern erklären. Um die Plausibilität der Erklärung für eine bestimmte Klassifikation zu überprüfen, wird das in der KI-Anwendung implementierte Modell auf einem Trainingsdatensatz mit randomisierten Labels neu trainiert und das Heatmap-Verfahren nun auf die Klassifikation des neu trainierten Modells angewandt. Stellt sich heraus, dass sich die Heatmaps bezüglich des ursprünglichen und des neu trainierten Modells stark ähneln, so liegt nahe, dass das Heatmap-Verfahren eher eine Art Kantendetektor ist, als dass es intrinsische Entscheidungsvorgänge der (sich stark unterscheidenden) Modelle erklären würde. Ferner sollte die Heatmap für Eingangsdaten weit außerhalb des Anwendungsbereichs, etwa einer nicht trainierten Klasse, idealerweise keine aussagekräftige Erklärung für irgendeine Klassifikation bezüglich der bekannten Klassen liefern.

[TR-R-EX-MA-04] Qualitätssicherung der Ergebnisse der Transparenzmethoden

Anforderungen: Do | Te

- Neben der Eignung für den gewählten Zweck sollten die Ergebnisse der ergriffenen introspektiven Maßnahmen den Ansprüchen aus **[TR-R-EX-KR-02]** genügen. Dies ist durch geeignete Tests zu verifizieren.
  - Die quantitativen Kriterien können durch statistische Tests überprüft werden, wie etwa in **[TR-R-NB-MA-05]**. Die verwendeten Testdatensätze sind zu dokumentieren.
  - Die qualitativen Kriterien können beispielsweise in Versuchen mit Testpersonen überprüft werden, wie etwa in **[TR-R-NB-MA-07]**.
- Durchführung und Ergebnisse der Tests werden dokumentiert, und es wird dargelegt, inwiefern die Testergebnisse die Erfüllung der in **[TR-R-EX-KR-02]** formulierten Kriterien belegen.

[TR-R-EX-MA-05] Vollständige Erfüllung eines Kriteriums

Anforderungen: Do | Te

- Falls mehrere Herangehensweisen/Transparenzmethoden auf die Erfüllung eines einzelnen Kriteriums ausgerichtet sind, d. h. für einen gemeinsamen Zweck zur Anwendung kommen, liegt eine Dokumentation vor, die begründet, dass das Kriterium durch die Kombination dieser Methoden tatsächlich vollständig erfüllt ist. Sofern dies technisch möglich ist, wird aktiv nach Lücken gesucht und das Ergebnis dokumentiert. Andernfalls kann begründet werden, warum eine solche Lücke nicht existieren kann. Diese Maßnahme findet nur dann Anwendung, wenn Herangehensweisen/Transparenzmethoden in Kombination genutzt werden.

46 Eine Orientierung bietet der Artikel: Adebayo, J. (November 2020). Sanity Checks for Saliency Maps. GitHub. https://github.com/adebayoj/sanity_checks_saliency (letzter Aufruf: 16.06.2021)

6.2.3.3 Einbettung

[TR-R-EX-MA-06] Visuell-interaktives Interface

Anforderungen: Do | Pr | Te

- Abhängig von der Komplexität der zugrundeliegenden KI-Komponente und der Daten kann es hilfreich und/oder notwendig sein, ein visuell-interaktives Interface einzusetzen, um etwa verschiedene Introspektionsmaßnahmen zu aggregieren und Daten sowie Metadaten visuell aufzubereiten. Insbesondere kann durch ein solches Interface die Masse an Daten für Expert*innen visuell heruntergebrochen und so das Datenverständnis erhöht werden. Ferner können Expert*innen semantische Hypothesen untersuchen, Cluster identifizieren und so bei der Suche nach Schwachstellen des Modells unterstützt werden (beispielsweise im Sinne eines *Closed Loop Testings*, vgl. **[VE-R-RE-MA-05]**).
  Beispiel: Bei der Untersuchung einer KI-Anwendung zur Fußgängererkennung fällt ein Bild auf, bei dem eine Person im roten Pullover nicht erkannt wird. Die Expert*innen möchten untersuchen, ob dieser Fehler bei ähnlichen Eingabedaten ebenfalls auftritt und somit eine systematische Schwachstelle darstellt. Innerhalb des visuellen Interfaces kann der entsprechende Bildbereich mit der Person im roten Pullover markiert werden. Ausgehend von diesem Beispielbild wird der Datensatz mithilfe einer Ähnlichkeitsmetrik nach ähnlichen Bildern durchsucht. Das Ergebnis ist ein gefilterter Datensatz, mithilfe dessen die Expert*innen nun die Performanz der KI-Anwendung für diesen speziellen Fall untersuchen können.
- Wurde ein visuell-interaktives Interface für die KI-Anwendung erstellt, so ist dieses in einer Dokumentation ausführlich zu beschreiben. Dabei ist unter anderem festzuhalten, welche Schnittstellen das visuell-interaktive Interface hat, welche Datenformate unterstützt werden, und wie die Einbettung der Ausgaben der KI-Anwendung bzw. des ML-Modells selbst umgesetzt ist. Weiter werden die implementierten Methoden zur visuellen Aufbereitung der Daten innerhalb des Interfaces beschrieben, z. B. welche Graphen, Tabellen, Bilddaten o. ä. dargestellt werden, welche Bereiche verknüpft sind und wie die Daten im Hintergrund verarbeitet werden. Die Interaktivität sowie die visuelle Darstellung bzw. Aggregation der Daten durch das visuell-interaktive Interface sollte sich an den Methoden der *Visual Analytics* orientieren. Außerdem wird die Bedienung des Interfaces anhand von Beispielen erläutert.
- Es wird dokumentiert, inwiefern das visuell-interaktive Interface zur Untersuchung von Schwachstellen des Modells verwendet wurde (beispielsweise im Sinne eines *Closed Loop Testings*, vgl. **[VE-R-RE-MA-05]**) und welche Schwachstellen dabei aufgedeckt wurden. Sofern dies mit Maßnahmen in der **Dimension: Verlässlichkeit (VE)** zusammenhängt, kann auf diese referenziert werden.
- Steht ein visuell-interaktives Interface zur Verfügung, das auch im Betrieb zur (systematischen) Schwachstellensuche des ML-Modells herangezogen werden soll, so ist hierzu ein Prozess zu etablieren. Insbesondere wird festgehalten, welche Aspekte bei der Analyse mithilfe des Interfaces fokussiert werden, nach welchen Kriterien Schwachstellen identifiziert werden, und welche Maßnahmen auf Erkenntnisse folgen.

[TR-R-EX-MA-07] Auswirkung der Einbettung

Anforderungen: Do | (Te)

- Es liegt eine Dokumentation vor, die zeigt, dass die Transparenzeigenschaften der KI-Anwendung nicht durch ihre Einbettung in ein Gesamtsystem verletzt werden. Sofern der begründete Verdacht besteht, dass dies nicht erfüllt sein könnte, wird dies anhand der Maßnahmen **[TR-R-EX-MA-04]** und ggf. **[TR-R-EX-MA-05]** für die KI-Komponente (schon eingebettet und als KI-Anwendung funktionstüchtig) getestet. Die Wahl der Testdatensätze, die Durchführung sowie die Testergebnisse sind zu dokumentieren.
  Beispiel: In einem selbstfahrenden Fahrzeug wird eine KI-Komponente in ein Online-Monitoring-System eingebettet, welches alle eingebetteten Module auf Basis komplexer Regeln überwacht. In bestimmten Situationen, zum Beispiel im externen Fehlerfall, kann die Einbettung die Ausgaben der Module überschreiben oder sogar abschalten, um in einen *Fail-Safe*-Modus überzugehen. Erratisches Verhalten dieses Online-Monitoring-Systems könnte sich demnach negativ auf die Nachvollziehbarkeit der Entscheidungen der KI-Komponente auswirken, insbesondere dann, wenn die komplexen Regeln eine Nachvollziehbarkeit der Handlungen der Einbettung erschweren.

- Dieser Test kann unter Umständen die vorherigen Tests (in **[TR-R-EX-MA-04]** und ggf. **[TR-R-EX-MA-05]**) auf der Ebene der KI-Komponente ersetzen, sofern dies ausreichend begründet wird.
- Wurden im Vorfeld aufgrund der Verletzung von Transparenzanforderungen Anpassungen an der Einbettung vorgenommen, so ist, falls möglich, die Änderungshistorie inklusive der Tests voriger Versionen anzugeben.

[TR-R-EX-MA-08] Beitrag der Einbettung

Anforderung: Do

- Abhängig von der Art der KI-Anwendung kann es zu einem gewissen Grad möglich sein, die Erfüllung von Transparenzanforderungen durch Eigenschaften der Einbettung zu erreichen. Es ist zu dokumentieren, inwiefern die Einbettung zur Erfüllung der Zielvorgaben bezüglich Transparenz beiträgt.
  Beispiel: Wenn Transparenz mit dem Ziel einer zusätzlichen Validierung und damit einhergehenden Verbesserung der KI-Komponente verbunden ist, so könnte dies eventuell auch durch das Sammeln zusätzlicher Trainingsdaten erreicht werden. Indem die Gesamtfunktion diejenigen Eingangsdaten, die mit hoher Unsicherheit oder fehlerhafter Prädiktion behandelt werden, an einen Server zurückspiegelt, können diese gesondert untersucht werden. In diesem Beispiel wäre das Vorgehen auch in der **Dimension: Datenschutz (DS)** und im **Risikogebiet: Einschätzung von Unsicherheit (UN)** der Dimension Verlässlichkeit zu berücksichtigen.

6.2.3.4 Maßnahmen für den Betrieb

Für diese Kategorie sind keine Maßnahmen vorgesehen.

6.2.4 Gesamtbewertung

[TR-R-EX-BW] Gesamtbewertung

Anforderung: Do

- Es wird unter Bezugnahme auf die zuvor beschriebenen Maßnahmen und Tests ausführlich dargelegt, dass die in **[TR-R-EX-KR-01]** und **[TR-R-EX-KR-02]** definierten Kriterien erfüllt sind.
- Sofern nicht alle in **[TR-R-EX-KR-01]** oder **[TR-R-EX-KR-02]** spezifizierten Anforderungen erfüllt werden, sind die Abweichungen zu dokumentieren. Dies gilt ebenfalls für nur teilerfüllte Anforderungen, bei denen etwa die Kriterien nicht, oder nicht immer, erreicht wurden.
- Es wird festgehalten, inwiefern in **[TR-R-EX-MA-02]** relevante negative Auswirkungen auf andere Dimensionen (insbesondere **Dimension: Verlässlichkeit (VE)**) festgestellt wurden, die in der dimensionsübergreifenden Gesamtbewertung aufgegriffen und abgewogen werden müssen.

6.3 Risikogebiet: Auditfähigkeit (AF)

Die Auditfähigkeit einer KI-Anwendung ist ein Oberbegriff für die detaillierte technische Dokumentation von Aufbau, Entwicklung und Funktionsweise der KI-Anwendung und insbesondere der dazu verwendeten Daten.

Auditfähigkeit einer KI-Anwendung kann verschiedenen Zwecken dienen. Zum einen können nachfolgende Prüfungen (interne oder externe Audits) mit ggf. abweichenden Fragestellungen erleichtert oder erst ermöglicht werden. Dokumentationen und geloggte Daten können beispielsweise eine essenzielle Voraussetzung sein, um in Haftungsfragen bestimmte Ausgaben der KI-Anwendung oder die Ursache von Fehlern nachzuverfolgen. Hierbei spielt auch die Reproduzierbarkeit von Ausgaben und des ML-Modells selbst eine Rolle. Zum anderen erleichtert eine ausführliche Dokumentation das Vornehmen von Änderungen bzw. Verbesserungen der KI-Anwendung.

Mögliche Anforderungen an die Auditfähigkeit können sich etwa aus dem **Risikogebiet: Transparenz für Expert*innen (EX)** ergeben. Spannungen hingegen können etwa mit der **Dimension: Datenschutz (DS)**, falls zu dokumentierende Daten personenbezogen sind, und der **Dimension: Sicherheit (SI)**, da Maßnahmen der Auditfähigkeit ggf. die Angreifbarkeit der KI-Anwendung erhöhen, auftreten und sind in der dimensionsübergreifenden Beurteilung zu diskutieren.

6.3.1 Risikoanalyse und Zielvorgaben

[TR-R-AF-RI-01] Risikoanalyse und Zielvorgaben

Anforderung: Do

- **Risikoanalyse:** Unter Beachtung des spezifischen Einsatzkontexts der KI-Anwendung sowie des rechtlichen Rahmens wird analysiert, welche Gefährdungen und potenziellen Schäden aufgrund von eingeschränkter Auditfähigkeit entstehen können. Dabei werden für jeden Aspekt der Auditfähigkeit (u. a. Dokumentation als Grundlage für externe Audits, Nachverfolgbarkeit und Reproduzierbarkeit) die Bedeutung für den vorliegenden Anwendungskontext sowie potenzielle Konsequenzen bei verschiedenen Graden der Nichterfüllung beleuchtet. Außerdem ist die Eintrittswahrscheinlichkeit und -häufigkeit der identifizierten potenziellen Schäden abzuschätzen.
- **Zielvorgaben:** Basierend auf den Erkenntnissen der Risikoanalyse werden qualitative Zielvorgaben bezüglich der verschiedenen Aspekte der Auditfähigkeit der KI-Anwendung definiert. Dabei wird ausführlich beschrieben, was Auditfähigkeit im Kontext der KI-Anwendung genau bedeutet, d. h. unter anderem, für welche Teile der KI-Anwendung, in welchem Umfang eine Dokumentation erforderlich ist (Trainingsdaten, Modelleigenschaften und Hyperparameter, Logging von Inputdaten, Ausgaben, Systemarchitektur, etc.) und in welchem Umfang die Ausgaben der KI-Anwendung nachverfolgbar oder gar reproduzierbar sein sollen.

6.3.2 Kriterien zur Zielerreichung

[TR-R-AF-KR-01] Grad der Auditfähigkeit der KI-Anwendung

Anforderung: Do

- Es werden Kriterien festgelegt und dokumentiert, anhand derer die Auditfähigkeit der KI-Anwendung sinnvoll beurteilt werden kann. Bei der Wahl der Kriterien sollten zumindest die folgenden Aspekte berücksichtigt werden:
  - Zugreifbarkeit, Quelle, Inhalt, Datum, Revisionsnummer sowie Möglichkeit, Notwendigkeit und Dauer der Speicherung von Trainings- und Testdaten
  - Vorliegen einer erläuternden Dokumentation der Systemarchitektur
  - Möglichkeit und Notwendigkeit des Reproduzierens von Ausgaben, ggf. innerhalb eines festgelegten Zeitraums, etwa durch Zugriff auf abgespeicherte Modellversionen

  - Möglichkeit und Notwendigkeit des Loggens von Entscheidungen und der hierzu notwendigen Daten (wie zum Beispiel Eingabedaten oder Zusatzinformationen, etwa über den Betriebszustand, die Einsatzumgebung oder *Random Seeds*, falls Ergebnisse nicht-deterministisch zustande kommen)

  Die Auflistung dient als Ausgangspunkt, um dem Anwendungskontext angemessene Kriterien festzulegen. Auch hier nicht aufgeführte Aspekte können zur Beurteilung der Auditfähigkeit herangezogen werden, wenn diese beschrieben und deren Wahl begründet werden.
- Zu jedem festgelegten Kriterium werden außerdem qualitative Zieleigenschaften oder ggf. quantitative Zielwerte dokumentiert, durch deren Erreichen die gemäß den Zielvorgaben in **[TR-R-AF-RI-01]** erforderliche Auditfähigkeit der KI-Anwendung hergestellt wird.
- Es ist zu begründen, dass die gewählten Kriterien und zugehörigen Zielwerte mit den in **[TR-R-AF-RI-01]** definierten Zielvorgaben konform sind.

6.3.3 Maßnahmen

6.3.3.1 Daten

[TR-R-AF-MA-01] Verfügbarkeit von Trainings- und Testdaten

Anforderungen: Do | Pr

- Es ist ein Speicherungskonzept zur Archivierung von Trainings- und Testdaten, die vor Inbetriebnahme der KI-Anwendung erhoben wurden, ausgearbeitet und dokumentiert. Gemäß dieses Konzepts stehen die Trainings- und Testdaten für Expert*innen und Entwickler*innen in einer Form zur Verfügung, sodass sie für notwendige Introspektionen verwendet werden können (vgl. **Risikogebiet: Transparenz für Expert*innen (EX)**). Die Zugriffsrechte sowie die Speicherung sind mit der **Dimension: Datenschutz (DS)** abzustimmen, falls unter die betreffenden Daten auch personenbezogene oder andere schützenswerte Daten fallen.
- Falls für die KI-Anwendung ein Neulernen in Frage kommt, so gibt es einen Prozess zur Protokollierung und Speicherung von Trainings- und Testdaten jedes Neulernens der KI-Anwendung. Insbesondere werden die betreffenden Trainings- und Testdaten versioniert und es ist stets nachvollziehbar, auf welchen Daten die aktuell im Betrieb befindliche Version trainiert wurde. Falls die KI-Anwendung kontinuierlich während des Betriebs weiterlernt, so ist auch **[TR-R-AF-MA-03]** zu berücksichtigen.

6.3.3.2 KI-Komponente

Für diese Kategorie sind keine Maßnahmen vorgesehen.

6.3.3.3 Einbettung

[TR-R-AF-MA-02] Software-Umgebung und Schnittstellen

Anforderung: Do

- Es liegt eine Dokumentation des Aufbaus der KI-Anwendung vor, die die verschiedenen Softwarekomponenten und weiteren Systemkomponenten (z. B. Cloud-Speicher) beschreibt und deren Wechselwirkung erläutert. Falls dies bereits im KI-Steckbrief oder an anderer Stelle beschrieben wurde, kann hierbei auf die entsprechende Dokumentation verwiesen werden.
- Es liegt eine Dokumentation der verwendeten Software-Bibliotheken und der genauen Versionen der verwendeten Packages vor.
- Es liegt eine Dokumentation der Einbettung der KI-Komponente mit den Beschreibungen aller Schnittstellen und der jeweiligen Eingabe- und Ausgabeformate an den Schnittstellen vor.

6.3.3.4 Maßnahmen für den Betrieb

Für Transparenzanforderungen zur Laufzeit kann es erforderlich sein, Eingangsdaten, Vorhersagen oder interne Zustände der KI-Anwendung zu protokollieren. Sollte eine derartige Anforderung bestehen, ist ein Prozess zur Protokollierung und Speicherung notwendiger Daten auszuarbeiten. Unter Umständen ist hierbei auch die **Dimension: Datenschutz (DS)** einzubeziehen. Die folgenden Maßnahmen greifen die Speicherung verschiedener Daten während des Betriebs auf.

[TR-R-AF-MA-03] Verfügbarkeit von Lerndaten aus dem Betrieb

Anforderung: Do

- Falls die KI-Anwendung während des Betriebs weiterlernt, oder Daten zur Validierung oder Erweiterung des Modells während des Betriebs erhoben werden, gibt es in Ergänzung zu **[TR-R-AF-MA-01]** ein Protokollierungs- und Speicherungskonzept für die dazu erhobenen Daten. Insbesondere dann, wenn das Modell im laufenden Betrieb ohne Revision anhand erhobener Daten weiter trainiert wird, sollten diese Daten den Expert*innen und Entwickler*innen auch im Nachhinein zur Verfügung stehen. Das Konzept sowie der Umfang des Loggings werden dokumentiert.
- Handelt es sich bei den zu speichernden Daten um personenbezogene Informationen oder schützenswerte Geschäftsdaten, so ist hierbei die **Dimension: Datenschutz (DS)** einzubeziehen.

[TR-R-AF-MA-04] Speicherung von Modell- und Trainingsparametern

Anforderung: Do

- Die Modellparameter, z. B. die Gewichte eines Neuronalen Netzes, werden gespeichert und versioniert.
- Die Hyperparameter, die die Trainingsprozedur bei einem Neulernen oder kontinuierlichen Weiterlernen des Modells während des Betriebs charakterisieren, werden gespeichert und versioniert.

[TR-R-AF-MA-05] Reproduzierbarkeit und Nachverfolgbarkeit

Anforderung: Do

- Es liegt ein Protokollierungs- und Speicherungskonzept für die Ausgaben der KI-Anwendung vor. Das Konzept sowie der Umfang des Loggings werden dokumentiert. Der Zugriff auf die geloggten Ausgaben (z. B. aufgrund von **[VE-R-RO-MA-08]** oder zur Korrektur der Ausgaben wie in **[AK-R-GE-MA-02]**) ist mit den entsprechenden Maßnahmen aus den anderen Dimensionen abzustimmen.
- Es liegt ein Protokollierungs- und Speicherungskonzept für die Eingaben der KI-Anwendung vor. Das Konzept sowie der Umfang des Loggings werden dokumentiert. Ggf. kann hierbei auf **[TR-R-AF-MA-03]** verwiesen werden. Der Zugriff auf die geloggten Eingaben (z. B. aufgrund von **[VE-R-RO-MA-07]**) ist mit den entsprechenden Maßnahmen aus den anderen Dimensionen abzustimmen.
- Falls sich Modellergebnisse nicht auf deterministische Art und Weise ergeben, und falls dies für die Reproduzierbarkeit oder Nachverfolgbarkeit erforderlich ist, werden die Zwischenrepräsentationen gespeichert, die bei der Inferenz des ML-Modells berechnet werden (z. B. Feature-Maps eines Neuronalen Netzes). Hierbei kann evaluiert werden, inwiefern das Loggen von Systemzuständen wie etwa *Random Seeds* ausreicht, um diese Zwischenrepräsentationen zuverlässig zu reproduzieren. Andernfalls muss begründet werden, warum auf die Speicherung dieser Daten verzichtet wird.
- Handelt es sich bei den zu speichernden Daten etwa um personenbezogene Information oder schützenswerte Geschäftsdaten, so ist hierbei die **Dimension: Datenschutz (DS)** einzubeziehen.
- Es ist darzulegen, inwiefern die ergriffenen Maßnahmen zu Herstellung der in **[TR-R-AF-KR-01]** geforderten Reproduzierbarkeit und Nachverfolgbarkeit beitragen. Falls der Umfang des Loggings und der Speicherung nicht ausreichen, um die Anforderungen zu erfüllen, ist dies zu begründen.

[TR-R-AF-MA-06] Logging von Benutzeranfragen

Anforderungen: Do | Pr

- Falls unter **[TR-R-NB-MA-08]** ein Prozess zur Beantwortung von Benutzeranfragen etabliert wurde, ist zu dokumentieren,
  - inwiefern Anfragen von Nutzer*innen und die darauf gegebenen Erklärungen geloggt werden.
  - wie lange diese Information gespeichert wird. Dabei ist der Schutz der Daten von Nutzer*innen und Betroffenen einzuhalten.
  - dass über die Speicherung der Anfragen Auskunft gegeben wird.
  - ob ein Prozess etabliert ist, um diese Anfragen unter Umständen zu löschen. Das Vorgehen ist ggf. mit der **Dimension: Datenschutz (DS)** abzustimmen.

6.3.4 Gesamtbewertung

[TR-R-AF-BW] Gesamtbewertung

Anforderung: Do

- Es liegt eine Dokumentation vor, in der erläutert wird, dass aufgrund der ergriffenen Maßnahmen die in **[TR-R-AF-KR-01]** definierten Kriterien zur Auditfähigkeit der KI-Anwendung erfüllt werden.
- Es wird festgehalten, inwiefern in diesem Risikogebiet relevante negative Auswirkungen auf andere Dimensionen (insbesondere **Dimension: Datenschutz (DS)**) festgestellt wurden, die in der dimensionsübergreifenden Beurteilung aufgegriffen und abgewogen werden müssen.
- Sofern nicht alle in **[TR-R-AF-KR-01]** spezifizierten Anforderungen erfüllt werden, sind die Abweichungen zu dokumentieren. Dies gilt ebenfalls für nur teilerfüllte Anforderungen, bei denen etwa die Kriterien nicht, oder nicht immer, erfüllt wurden.

6.4 Risikogebiet: Beherrschung der Dynamik (BD)

Das Risikogebiet Beherrschung der Dynamik soll sicherstellen, dass die Transparenz der KI-Anwendung aufrechterhalten wird.

Zum einen wird das Risiko behandelt, dass im Laufe des Betriebs etablierte Transparenzeigenschaften der KI-Anwendung verloren gehen können, falls sich etwa das zu erklärende ML-Modell im Sinne eines *Model Drifts* ändert. Beispielsweise kann eine Änderung des ML-Modells dazu führen, dass die Stabilität von Erklärungen nachlässt. Zum anderen besteht das Risiko, dass sich die Anforderungen an die Transparenz der KI-Anwendung verändern, beispielsweise aufgrund externer Faktoren wie etwa Bedürfnissen von Nutzer*innen oder neuen Gesetzen. Zusätzlich kann eine Weiterentwicklung im State of the Art dazu führen, dass bisher nicht erreichbare Anforderungen realisierbar oder andere obsolet werden. Eine abstrakte Anforderung an Transparenz könnte darin bestehen, diese Entwicklungen zu überwachen und aufzugreifen.

6.4.1 Risikoanalyse und Zielvorgaben

[TR-R-BD-RI-01] Risikoanalyse und Zielvorgaben

Anforderung: Do

- **Risikoanalyse:** Es wird analysiert, welche (externen) Faktoren bzw. Umstände einen Einfluss auf die in den vorangegangenen Risikogebieten etablierten Transparenzanforderungen haben. Darauf aufbauend wird abgeschätzt, mit welcher Wahrscheinlichkeit sich die Anforderungen an die Transparenz der KI-Anwendung absehbar ändern werden, und welche Auswirkungen bzw. potenzielle Schäden daraus resultieren können. Zudem werden mögliche (externe oder in der KI-Anwendung selbst liegende) Ursachen für den Verlust bestehender Transparenzeigenschaften der KI-Anwendung während des Betriebs identifiziert und deren Eintrittswahrscheinlichkeit abgeschätzt. Ferner wird untersucht, welche Schäden entstehen können, wenn solch ein Verlust eintritt.
- **Zielvorgaben:** Basierend auf der Risikoanalyse werden qualitative Zielvorgaben für den Betrieb formuliert, die umreißen, in welchem Umfang und mit welchem Vorgehen die Aufrechterhaltung bzw., falls erforderlich, Anpassung der etablierten Transparenzeigenschaften während des Betriebs gewährleistet werden soll, damit Risiken auf ein vertretbares Maß gesenkt werden.

6.4.2 Kriterien zur Zielerreichung

[TR-R-BD-KR-01] Überprüfung und Anpassung von Transparenzanforderungen

Anforderung: Do

- Basierend auf der Risikoanalyse **[TR-R-BD-RI-01]** werden externe Faktoren festgelegt, die aufgrund ihrer möglichen Auswirkungen auf die Transparenzanforderungen, die in den vorangegangenen Risikogebieten etabliert wurden, beobachtet werden sollen. Entsprechend der geschätzten Änderungsrate dieser Faktoren ist das dafür vorgesehene Prüf- bzw. Beobachtungsintervall zu definieren und dokumentieren.
- In Bezug auf die Änderung relevanter externer Faktoren wird ein Schwellwert bzw. das qualitative Ausmaß der Abweichung festgelegt und dokumentiert, ab der eine Anpassung der bisher etablierten Transparenzanforderungen (die entsprechend **[TR-R-BD-KR-02]** regelmäßig überprüft werden) initiiert wird.
- Es ist darzulegen, dass die gesetzten Kriterien mit den Zielvorgaben aus **[TR-R-BD-RI-01]** konform sind.

[TR-R-BD-KR-02] Aufrechterhaltung von Transparenzeigenschaften

Anforderung: Do

- Es werden Kriterien an einen Prozess zur regelmäßigen Überprüfung von Transparenzeigenschaften gegen die bestehenden Anforderungen festgelegt und dokumentiert. Bei der Wahl der Kriterien zur Beurteilung des Prozesses sollten mindestens die folgenden Punkte (quantitativ oder qualitativ) berücksichtigt werden:
  - Prüfintervall. Dabei wird aufgezeigt, inwieweit die Prüfungen der Transparenzeigenschaften auch von anderen, regelmäßigen Prüfungen angestoßen werden. Diese Vorgehensweise bietet sich an, wenn eine erneute Evaluierung der Transparenzeigenschaften nicht nur aus direkten Änderungen der Transparenzanforderungen, sondern aus anderweitig initiierten Änderungen der KI-Anwendung, wie etwa einem Neutraining, resultieren.
  - Umfang und Art der Überprüfung bzw. verwendete Methoden
  - Schwellwert bzw. qualitatives Ausmaß der Abweichung von den Anforderungen, ab der korrektive Maßnahmen ergriffen werden

  Zu jedem Kriterium sind quantitative Zielwerte bzw. qualitative Zieleigenschaften zu spezifizieren.
- Es ist darzulegen, dass die gesetzten Kriterien die Zielvorgaben aus **[TR-R-BD-RI-01]** angemessen abbilden.

6.4.3 Maßnahmen

6.4.3.1 Daten

Für diese Kategorie sind keine Maßnahmen vorgesehen.

6.4.3.2 KI-Komponente

Für diese Kategorie sind keine Maßnahmen vorgesehen.

6.4.3.3 Einbettung

Für diese Kategorie sind keine Maßnahmen vorgesehen.

6.4.3.4 Maßnahmen für den Betrieb

[TR-R-BD-MA-01] Beobachtung externer Faktoren

Anforderungen: Do | Pr

- Es wird ein Monitoring-Prozess bezüglich externer Faktoren etabliert, die die in den vorangegangenen Risikogebieten festgestellten Transparenzanforderungen beeinflussen können. Dabei sind mindestens die in **[TR-R-BD-KR-01]** genannten Faktoren zu berücksichtigen. Abhängig von der Art der Transparenzanforderungen kann der Prozess auf Methoden wie Beobachtung und Analyse basieren, aber er kann beispielsweise auch das Erheben von Nutzerfeedback durch Fragebögen beinhalten. Art und Umfang des Prozesses sind zu dokumentieren und es ist ausführlich zu begründen, inwiefern der Prozess zur Erfüllung der Zielvorgaben beiträgt.
- Außerdem wird dokumentiert, wie im Rahmen des Prozesses eine Anpassung der bisher etablierten Transparenzanforderungen gemäß **[TR-R-BD-KR-01]** vorgenommen wird.
- Weiterhin ist zu beschreiben, welche Schritte auf eine Anpassung der Transparenzanforderungen folgen. Insbesondere wird dargelegt, inwiefern Anpassungen an der KI-Anwendung initiiert oder vorgenommen werden (vgl. **[TR-R-BD-MA-02]**).

[TR-R-BD-MA-02] Überprüfung und Aufrechterhaltung der Transparenzeigenschaften

Anforderungen: Do | Pr | Te

- Es wird ein Prozess etabliert, der gemäß den Vorgaben in **[TR-R-BD-KR-02]** die Aufrechterhaltung der bestehenden Transparenzeigenschaften der KI-Anwendung überprüft und, falls erforderlich, Korrekturen initiiert bzw. vornimmt.
  - Art und Umfang der pro Prüfintervall durchgeführten Tests werden dokumentiert. Hierbei kann ggf. auf die Methodik in **[TR-R-NB-MA-05]**, **[TR-R-NB-MA-07]** und **[TR-R-EX-MA-04]** zurückgegriffen werden. Insbesondere ist festzuhalten, welche Testdatensätze verwendet werden, und deren Wahl zu begründen. Sofern Testdatensätze im Betrieb erhoben werden, sind diese im Prüfprozess gegenüber den vor Inbetriebnahme verwendeten Testdaten vorzuziehen (angemessener Umgang mit *Concept Drift*). Andernfalls muss eine Vernachlässigung dieser Testdatensätze begründet werden.
  - Es wird beschrieben, wie mit den Dokumentationen aus den Prüfintervallen und insbesondere den Testergebnissen weiter verfahren wird. Dabei ist zu beschreiben, welche Schritte zur Wiederherstellung der Transparenzeigenschaften eingeleitet werden, falls bei einer Überprüfung unzureichende Testergebnisse erzielt werden.

6.4.4 Gesamtbewertung

[TR-R-BD-BW] Gesamtbewertung

Anforderung: Do

- Es wird dargelegt, dass Prozesse zur Beobachtung externer Faktoren sowie zur regelmäßigen Überprüfung der KI-Anwendung aufgesetzt wurden, die die Anforderungen in **[TR-R-BD-KR-01]** und **[TR-R-BD-KR-02]** erfüllen.
- Sofern nicht alle in **[TR-R-BD-KR-01]** und **[TR-R-BD-KR-02]** spezifizierten Anforderungen erfüllt werden, sind die Abweichungen zu dokumentieren. Dies gilt ebenfalls für nur teilerfüllte Anforderungen, bei denen etwa die Kriterien nicht, oder nicht immer, erreicht wurden.

Zusammenfassende Betrachtung

[TR-Z] Zusammenfassende Betrachtung der Dimension

Anforderung: Do

- Falls für diese Dimension ein mittlerer oder hoher Schutzbedarf besteht, ist eine Dokumentation über die verbleibenden Restrisiken zu erstellen. Zunächst werden die Restrisiken aus den verschiedenen Risikogebieten dieser Dimension zusammengefasst. Anschließend wird unter Berücksichtigung des Schutzbedarfs analysiert, ob die identifizierten Restrisiken insgesamt als vernachlässigbar, nicht vernachlässigbar (aber vertretbar) oder unvertretbar zu bewerten sind. Das Ergebnis der Analyse ist zu erläutern.
- Falls potenziell negative Auswirkungen von Risiken oder Maßnahmen dieser Dimension auf andere Dimensionen, beispielsweise Verlässlichkeit oder Sicherheit, festgestellt wurden, sind diese zu dokumentieren.
- Es wird ein Fazit über die Dimension gezogen, welches insbesondere die Bewertung der Restrisiken enthält.

7. Dimension: Verlässlichkeit (VE)

Beschreibung und Zielsetzung

Aus technischer Sicht stellt die Verlässlichkeit einer KI-Anwendung einen Sammelbegriff dar, der unterschiedliche Aspekte der Güte ihrer KI-Komponente umfasst: Die Korrektheit der Ausgaben, die Einschätzung der ML-Modellunsicherheit, die Robustheit gegenüber gestörten oder manipulierten Eingaben sowie unerwarteten Situationen, und nicht zuletzt das Abfangen von Fehlern.[47, 48]

Profundes Anwendungswissen ist nötig, um diese verschiedenen Aspekte der Verlässlichkeit zu bewerten und festzulegen, unter welchen Voraussetzungen die KI-Anwendung als verlässlich einzustufen ist. Die Übersetzung der Anforderungen in quantitative Maße und Zielwerte erfordert Domänenwissen sowie mathematisch-technische Expertise und ist naturgemäß niemals vollständig. Gleiches gilt für die Beschreibung des Anwendungsbereichs der KI-Anwendung. Als Anwendungsbereich werden diejenigen im Betrieb zu erwartenden Eingabedaten bezeichnet, für die eine korrekte Verarbeitung durch die KI-Komponente vorgesehen ist.

Im **Risikogebiet: Verlässlichkeit im Regelfall (RE)** wird dieser Anwendungsbereich möglichst genau spezifiziert und formalisiert, damit sichergestellt werden kann, dass die verwendeten Trainings- und Testdaten den Anwendungsbereich angemessen abdecken. Meist sind hierbei auch Störungen der Eingabedaten, die im regulären Betrieb auftreten können und zum Anwendungsbereich zählen, intrinsisch in der Datenbasis abgebildet. Gleichzeitig kann es für manchen Einsatzkontext sinnvoll sein, derartige Störungen herauszugreifen und gezielt zu behandeln, um die Abdeckung der Daten sowie die Performanz der KI-Anwendung zu stärken. Dies erfolgt im **Risikogebiet: Robustheit (RO)**, dessen Ziel es ist, eine möglichst konsistente Performanz der KI-Komponente auch an der Grenze des Anwendungsbereichs zu gewährleisten. Dazu werden störungsbehaftete oder manipulierte Eingabedaten wie etwa Sensorrauschen oder adversariale Beispiele adressiert.

Zur Verlässlichkeit einer KI-Anwendung gehört darüber hinaus der angemessene Umgang mit Abweichungen vom Regelfall. Insbesondere bei direkter Mensch-Maschine-Interaktion können neuartige, weil für Menschen unerwartete oder unbekannte Fehlermodi, zu potenziell kritischen, da nicht eingeübten, Situationen führen. Daher sollten Eingabedaten, die weit außerhalb des definierten Anwendungsbereichs liegen und für die dementsprechend keine korrekte Verarbeitung durch die KI-Komponente zu erwarten ist, abgefangen werden. Im **Risikogebiet: Abfangen von Fehlern auf Modellebene (AF)** wird die Verlässlichkeit der KI-Anwendung dahingehend untersucht, inwiefern potenziellen Fehlern bereits durch Detektionsstrategien auf Ebene der KI-Komponente vorgebeugt wird. Dieses Risikogebiet ergänzt die klassischen Methoden zur Fehlertoleranz und zum *Fail-Safe* auf Ebene der Einbettung, wie sie im **Risikogebiet: Funktionale Sicherheit (FS)** in der Dimension Sicherheit behandelt werden.

---

47 Die Darstellung in diesem Abschnitt sowie teils in den folgenden Abschnitten ist stark angelehnt an das Kapitel »3.4 Verlässlichkeit« des Whitepapers: Poretschkin, M., et al. (2019). Vertrauenswürdiger Einsatz von Künstlicher Intelligenz. Sankt Augustin: Fraunhofer-Institut für Intelligente Analyse und Informationssysteme IAIS. https://www.ki.nrw/wp-content/uploads/2020/03/Whitepaper_KI-Zertifizierung.pdf (letzter Aufruf: 18.06.2021)

48 Eine breit aufgestellte technische Übersicht über hier diskutierte Bereiche und Methoden findet sich beispielsweise in: Houben, S. et al. (2021). Inspect, understand, overcome: A survey of practical methods for ai safety. https://www.ki-absicherung-projekt.de/fileadmin/KI_Absicherung/Downloads/KI-A_20201221_Houben_et_al_-_Inspect__Understand__Overcome.pdf (letzter Aufruf: 18.06.2021)

Die Verlässlichkeit einer KI-Anwendung kann außerdem in verschiedener Hinsicht durch eine realistische Unsicherheitsbewertung gesteigert werden. Zum einen bietet die Einschätzung von Unsicherheit die Möglichkeit, Schwächen des ML-Modells aufzudecken und darauf zu reagieren. Zum anderen kann eine Unsicherheitsbewertung als Indikator dienen, falls Eingaben an der Grenze oder außerhalb des Anwendungsbereichs liegen. Dadurch können insbesondere potenzielle Fehler der KI-Anwendung abgefangen werden. Eine korrekte Unsicherheitsschätzung ist eine notwendige Voraussetzung, um diese Vorteile auszuschöpfen. Damit einhergehende Anforderungen werden im **Risikogebiet: Einschätzung von Unsicherheit (UN)** untersucht.

Die Risikogebiete der Dimension Verlässlichkeit sind die folgenden:

1. **Verlässlichkeit im Regelfall:** Dieses Risikogebiet behandelt das Risiko fehlerhafter Vorhersagen durch die KI-Komponente auf regulären Eingabedaten[49].

2. **Robustheit:** Dieses Risikogebiet adressiert Risiken, die sich bei störungsbehafteten oder manipulierten Eingabedaten ergeben, für die jedoch eine korrekte Verarbeitung durch die KI-Komponente beabsichtigt ist. Dabei werden sowohl qualitative als auch quantitative Abweichungen der Eingabedaten betrachtet, wie etwa Rauschen oder adversariale Beispiele.

3. **Abfangen von Fehlern auf Modellebene:** Dieses Risikogebiet behandelt Risiken aus Eingabedaten, die nicht im Anwendungsbereich liegen und für die eine korrekte Bearbeitung durch die KI-Komponente nicht zu erwarten ist. Diese sollen durch eine Detektionsstrategie abgefangen werden.

4. **Einschätzung von Unsicherheit:** Dieses Risikogebiet betrachtet Risiken, die sich daraus ergeben, dass keine Unsicherheitsschätzung bezüglich Ausgaben stattfindet, oder, dass diese unrealistisch bzw. unbrauchbar ist.

5. **Beherrschung der Dynamik:** Dieses Risikogebiet befasst sich mit dem Risiko, dass das in der KI-Komponente implementierte ML-Modell aufgrund von unbeabsichtigten *Model Drifts* oder Veränderungen des Anwendungskontexts (*Concept Drift*) Einbußen an Performanz oder hinsichtlich anderer Anforderungen erfährt.

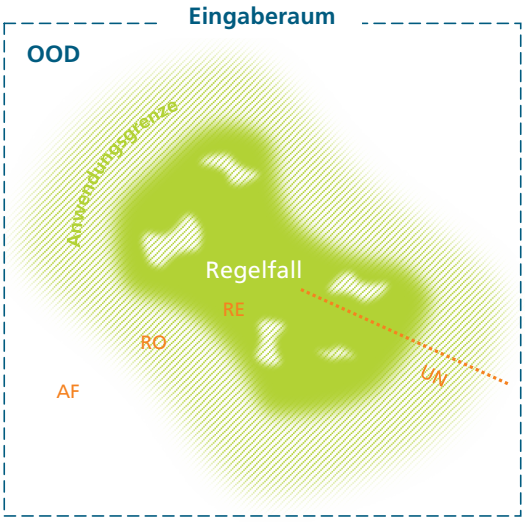

Abbildung 7: Darstellung der verschiedenen Gebiete des Eingaberaums und Zuordnung zu den Risikogebieten der Dimension Verlässlichkeit.

49 Als reguläre Eingabedaten werden im Folgenden solche Eingabedaten bezeichnet, die dem Anwendungsbereich entstammen, d. h. für die eine korrekte Verarbeitung durch die KI-Anwendung beabsichtigt ist.

Schutzbedarfsanalyse

Der Schutzbedarf der Dimension Verlässlichkeit ist durch die maximale Höhe des Schadens bestimmt, den (fehlerhafte) Ausgaben der KI-Anwendung ohne bewusste nachträgliche menschliche Revision verursachen können. Beispielsweise besteht ein tendenziell hoher Schutzbedarf für KI-Anwendung, die in autonome Robotik-Systeme eingebettet sind, oder in Entscheidungssysteme, die über die Verteilung von Gütern oder etwa über die medizinische Behandlung von Personen bestimmen.

Im Gegensatz zu den anderen Dimensionen ist für Verlässlichkeit ein *geringer* Schutzbedarf ausgeschlossen. Dies rührt aus der Tatsache, dass ein geringer Schutzbedarf in Hinblick auf Verlässlichkeit bedeutet, dass die Ausgaben der KI-Anwendung bei fehlerhafter Funktion keinen oder nur einen geringen Schaden erzeugen können. Insbesondere würden bei fehlerhaften Ausgaben keine Personenschäden entstehen und potenzielle finanzielle Schäden, etwa durch Verdienstausfall oder Rufschädigung, gar nicht oder nur in geringem Ausmaß drohen. Anders gesagt, die KI-Anwendung wäre in diesem Fall völlig unkritisch und könnte gleichermaßen durch ein nicht-KI-basiertes System, das zufällige Ausgaben erzeugt, ersetzt werden. Aus diesem Grund scheint die Prüfung einer KI-Anwendung, die bezüglich der Verlässlichkeit *geringen* Schutzbedarf hat, nicht erforderlich, sodass dieser Fall hier ausgeschlossen wird.

Der Schutzbedarf wird folgendermaßen kategorisiert:

Hoch	Eine Fehleinschätzung der KI-Anwendung kann zu Personenschaden oder hohem finanziellen Schaden führen. **Beispiele:** Fußgängererkennung im autonomen Fahrzeug, automatisierte Kreditvergabeentscheidungen, Empfehlungen medizinischer Behandlungen
Mittel	Eine Fehleinschätzung der KI-Anwendung kann maximal zu mittlerem finanziellen Schaden führen. **Beispiele:** Routenplaner geringer Qualität bedingt erhöhten Energieverbrauch und Zeitaufwand, fehlerhafte Vorhersage der Maschinenauslastung in einer Fertigung ruft Verzögerungen hervor, defekte Hinderniserkennung eines Staubsaugerroboters führt zu Sachschäden

Anmerkung: Die Schadenshöhe hängt in allen genannten Beispielen von den jeweiligen Umständen ab und muss nicht zwingend in die hier gewählte Kategorie fallen.

[VE-S] Dokumentation der Schutzbedarfsanalyse

Anforderung: Do

- Der Schutzbedarf der KI-Anwendung für die Dimension Verlässlichkeit wird als *mittel* oder *hoch* festgelegt. Die Wahl der Kategorie (*mittel oder hoch)* wird unter Bezugnahme auf die obige Tabelle ausführlich begründet.

In jedem Fall (*mittlerer* oder *hoher* Schutzbedarf) ist im Folgenden jedes Risikogebiet im Detail zu behandeln.

7.1 Risikogebiet: Verlässlichkeit im Regelfall (RE)

Das Risikogebiet Verlässlichkeit im Regelfall trägt allen Fehleinschätzungen einer KI-Anwendung Rechnung, die auf Eingabedaten aus dem Anwendungsbereich potenziell auftreten und zu Schäden führen können. Die Kriterien und Maßnahmen in diesem Risikogebiet adressieren das Gesamtrisiko einer Fehleinschätzung der KI-Komponente, mit dem Ziel, dieses auf ein im Kontext der Anwendung vertretbares Maß zu beschränken. Zusätzlich zu der universellen Betrachtung in diesem Risikogebiet werden spezifische Aspekte zur Vermeidung von Fehleinschätzungen bzw. Fehlern im Anwendungsbereich im **Risikogebiet: Robustheit (RO)** und im **Risikogebiet: Einschätzung von Unsicherheit (UN)** herausgegriffen und tiefergehend behandelt.

Ein erster wichtiger Schritt in der Auseinandersetzung mit dem Regelfall besteht darin, den Anwendungsbereich zu spezifizieren. Um eine angemessene Qualität der KI-Anwendung auf den im Betrieb zu erwartenden Eingabedaten zu gewährleisten, sollte sichergestellt sein, dass der Anwendungsbereich durch die Trainingsdaten ausreichend abgedeckt ist. Die korrekte Implementierung der Trainingsroutinen und des fertig trainierten Modells stellen eine weitere notwendige Voraussetzung für die Qualität der KI-Anwendung dar.[50]

Die Anforderungen an die Verlässlichkeit im Regelfall ergeben sich maßgeblich aus dem Einsatzkontext der KI-Anwendung. Um als objektiv überprüfbares Kriterium dienen zu können, müssen bestehende Anforderungen an die KI-Anwendung zunächst quantifiziert werden. Insbesondere im Fall qualitativer Anforderungen wie etwa »Kundenzufriedenheit« oder »Entlastung von Mitarbeiter*innen« ist die Übersetzung in quantitative Metriken und Zielintervalle, gegen die in Tests geprüft werden kann, immer verlustbehaftet. Um ein *Overfitting* auf eine bestimmte Metrik, d. h. eine einseitige Optimierung der KI-Komponente in Hinblick auf einen einzelnen Zielaspekt, zu vermeiden, sollten zur Beurteilung der Verlässlichkeit möglichst verschiedene Performanz-Metriken betrachtet werden. Zur Beurteilung, ob die Zielintervalle eingehalten werden, sollten Tests etabliert werden, die auf die betreffende KI-Anwendung und insbesondere deren Einsatzkontext abgestimmt sind.

7.1.1 Risikoanalyse und Zielvorgaben

Zur Durchführung der Risikoanalyse für das Risikogebiet Verlässlichkeit im Regelfall muss in einem ersten Schritt der (zulässige) Anwendungsbereich spezifiziert werden. Hierauf aufbauend erfolgt eine umfängliche Abschätzung von Art und Höhe möglicher Schäden, die durch die KI-Anwendung bei anwendungskonformer Nutzung direkt oder indirekt verursacht werden können.

[VE-R-RE-RI-01] Festlegen des Anwendungsbereichs und Risikoabschätzung

Anforderung: Do

Es liegt eine Dokumentation vor, die folgende Punkte qualitativ und konzeptionell adressiert:

- **Anwendungsbereich:** Als Anwendungsbereich wird der Teil der im Betrieb zu erwartenden Eingabedaten bezeichnet, für den eine korrekte Verarbeitung durch die KI-Komponente beabsichtigt ist. Dieser ist ausführlich zu beschreiben. Außerdem ist der Anwendungsbereich anhand von Beispielen grob abzugrenzen (störungsbehaftete Eingabedaten bzw. Grenzfälle des Anwendungsbereichs, unter anderem nicht-semantische Störungen wie etwa Rauschen, werden im **Risikogebiet: Robustheit (RO)** genauer spezifiziert).
  Wenn möglich, ist der hier definierte Anwendungsbereich gegenüber verwandten oder vermeintlich ähnlichen Anwendungsbereichen abzugrenzen, für die die KI-Anwendung nicht entwickelt wurde und bzgl. derer ihre Verlässlichkeit daher nicht geprüft wird.
- **Risikoanalyse**: Es liegt eine Abschätzung darüber vor, welche Risiken durch eine zu geringe Verlässlichkeit der KI-Anwendung bei regulären Eingabedaten bestehen. Außerdem wird beschrieben, welche potenziellen

50 Die Darstellung in diesem Abschnitt ist stark angelehnt an das Kapitel »3.4 Verlässlichkeit« des Whitepapers: Poretschkin, M., et al. (2019). Vertrauenswürdiger Einsatz von Künstlicher Intelligenz. Sankt Augustin: Fraunhofer-Institut für Intelligente Analyse und Informationssysteme IAIS. https://www.ki.nrw/wp-content/uploads/2020/03/Whitepaper_KI-Zertifizierung.pdf (letzter Aufruf: 18.06.2021)

Schäden hieraus resultieren und, falls möglich, mit welcher Häufigkeit diese auftreten können. Hierbei wird für jedes identifizierte Risiko abgeschätzt, ob der damit verbundene potenzielle Schaden vertretbar ist oder nicht.
- **Zielvorgaben**: Es wird dokumentiert, welche Restrisiken für den vorliegenden Einsatzkontext vertretbar sind. Ferner werden Zielvorgaben für Entwicklung und Betrieb gemacht, sodass bei deren Erfüllung das tolerierbare Restrisiko nicht überschritten wird.

7.1.2 Kriterien zur Zielerreichung

Im Folgenden werden Methoden und Metriken zur Risikoabschätzung sowie zur Konkretisierung des Anwendungsbereichs festgelegt. Dabei werden für alle herangezogenen Metriken Wertebereiche (Zielintervalle) bestimmt, die das in den Zielvorgaben angestrebte, vertretbare Risiko quantifizieren. Anhand dieser Zielintervalle kann die Verlässlichkeit der KI-Anwendung in der Gesamtbewertung beurteilt werden.

[VE-R-RE-KR-01] Quantifizierung der Verlässlichkeit

Die Verlässlichkeit einer KI-Anwendung wird in mathematisch-statistischen Maßen bzw. Metriken festgehalten. Dabei ist zwischen allgemeinen Performanz-Metriken und der Loss-Funktion zu unterscheiden.

Letztere ist maßgeblicher Bestandteil des Trainings vieler ML-Modelle. Insbesondere im Fall überwachter Maschineller Lernverfahren wird in der Regel das Modell mit Lernalgorithmen wie etwa *Stochastic Gradient Descent* in jedem Trainingsschritt so angepasst, dass die Loss-Funktion auf dem Trainingsdatensatz schrittweise optimiert wird. Der Wert der Loss-Funktion (auch Loss genannt) auf den Trainingsdaten, verglichen mit dem Loss auf den Testdaten, gibt Aufschluss über die Qualität des Trainings und kann z. B. auf mögliches *Overfitting* hindeuten.

Eine Performanz-Metrik hingegen ist der allgemeine Begriff für ein Maß bzw. eine Metrik, die beurteilt, wie gut/schlecht ein Modell bzw. die darauf basierende KI-Anwendung eine gestellte Aufgabe löst. Die Performanz-Metrik zur Beurteilung der Qualität eines Modells kann sich von seiner Loss-Funktion (falls vorhanden) unterscheiden. Insbesondere ist sie nicht Teil des Trainingsprozesses, sondern wird auf Validierungs- oder Testdaten berechnet, um die Performanz der KI-Komponente in Bezug auf eine bestimmte Aufgabe zu messen.

Für Regression und Klassifikation, zwei zentrale Problemstellungen des überwachten Maschinellen Lernens, sind im Folgenden gängige Loss-Funktionen aufgelistet. Diese Loss-Funktionen werden in der Praxis oftmals durch einen Regularisierungsterm für die Modellparameter (z. B. die Gewichte eines Neuronalen Netzes) ergänzt, um *Overfitting* entgegenzuwirken.

Regression:[51]
- *Squared Error Loss*
- *Absolute Error Loss*

Classification:[52]
- *Cross-Entropy Loss*
- *Hinge Loss*
- *Kullback-Leibler Divergence*
- *Brier Score*[53]

---

51 Details zu den Loss-Funktionen finden sich in: Patterson, J., & Gibson, A. (2017). Deep Learning - A Practitioner's Approach. O'Reilly Media; so wie teilweise auch in: Goodfellow, I., Bengio, Y., Courville, A. (2016). Deep Learning. MIT Press.

52 Details zu den Loss-Funktionen (mit Ausnahme des Brier Scores) finden sich in: Patterson, J., & Gibson, A. (2017). Deep Learning - A Practitioner's Approach. O'Reilly Media; so wie teilweise auch in: Goodfellow, I., Bengio, Y., Courville, A. (2016). Deep Learning. MIT Press.

53 Siehe: Brier, G. W. (1950). Verification of Forecasts Expressed in Terms of Probability, *Monthly Weather Review*, *78*(1), 1-3. Retrieved Jun 23, 2021. https://journals.ametsoc.org/view/journals/mwre/78/1/1520-0493_1950_078_0001_vofeit_2_0_co_2.xml (letzter Aufruf: 23.06.2021).

Die folgende Auflistung gibt einen Überblick über wichtige Performanz-Metriken für verschiedene (auch unüberwachte) Problemstellungen des Maschinellen Lernens. Es gibt eine Vielzahl weiterer Metriken.

Regression:[54]
- *(Mean) Squared Error*
- *(Mean) Absolute Error*

Classification:[55]
- *(Mean) Accuracy*
- *F1-Score*
- *Precision and Recall*
- *Sensitivity and Specificity*
- *AUC-Value*[56]

Ranking:
- *Mean Reciprocal Rank*[57]
- *Discounted Cumulative Gain*[58]

Clustering:
- *Silhouette Value*[59]
- *Adjusted Mutual Information Score*[60]
- *Completeness Score*[61]

Computer Vision:[62]
- *Peak Signal-to-Noise Ratio (SNR)*
- *Structural Similarity Index*[63]
- *(Mean) Intersection over Union (mIoU)*

Natural Language Processing:
- *Perplexity Score*[64]
- *BLEU Score*[65]

---

54 Referenzen wie in Fußnote 50

55 Referenzen wie in Fußnote 50 (außer *AUC-value*)

56 Siehe: Smola, A., Vishwanathan, S. V. N. (2008). Introduction to Machine Learning, Cambridge University Press.

57 Siehe: Radev, D. R. et al. (2002). Evaluating Web-based Question Answering Systems. Proceedings of the Third International Conference on Language Resources and Evaluation, pp. 1153-1156.

58 Vergleiche beispielsweise: Pedregosa, F. et al. (2020). Scikit-Learn: User Guide (Release 0.23.2). https://scikit-learn.org/stable/user_guide.html (letzter Aufruf: 23.06.2021).

59 Andrienko, N. et al. (2020). Visual Analytics for Data Scientists. Springer Nature Switzerland.

60 Referenz siehe Fußnote 56

61 Siehe: Bonaccorso, G. (2018). Mastering Machine Learning Algorithms. Packt Publishing.

62 Für *Peak SNR* sowie *mIoU*: Szeliski, R. (2021). Computer Vision: Algorithms and Applications (2nd Edition). Springer.

63 Vergleiche: Dosselmann, R., Yang, X.D. A comprehensive assessment of the structural similarity index. *SIViP* 5, 81–91 (2011). https://doi.org/10.1007/s11760-009-0144-1 (letzter Aufruf: 23.06.2021).

64 Beispielsweise in: Buduma, N. (2017). Fundamentals of Deep Learning. O'Reilly Media

65 Aus: Papineni, K. et al. (2002). BLEU: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL '02). Association for Computational Linguistics, USA, 311–318. https://doi.org/10.3115/1073083.1073135 (letzter Aufruf: 23.06.2021).

Anforderung: Do

- Es wird dokumentiert, welche Performanz-Metrik(en) zur Beurteilung der Verlässlichkeit der KI-Anwendung im Anwendungsbereich herangezogen werden soll(en). Dabei ist zu begründen, warum die gewählten Metriken zur Bewertung der Aufgabenerfüllung der KI-Anwendung angemessen sind und die spezifischen Anforderungen an die Qualität des Modells, die beispielsweise aus dem Business-Kontext der KI-Anwendung hervorgehen, ausreichend abbilden. Falls das Modell mit einer Loss-Funktion trainiert wird, so ist auch deren Wahl anzugeben und zu begründen. Wenn keine der in der obigen Auflistung angegebenen Evaluationsgrößen verwendet wird, so ist dies zu begründen. Darüber hinaus muss in diesem Fall dokumentiert sein, weshalb die stattdessen gewählte Loss-Funktion oder Metrik sinnvoll und für die KI-Anwendung passend ist.
- Ferner werden Zielintervalle für die gewählten Performanz-Metrik(en) und ggf. die Loss-Funktion im Anwendungsbereich festgelegt. Die Wahl der zu erreichenden Werte wird angesichts des vorliegenden Einsatzkontextes und der Kritikalität der durch die KI-Anwendung zu lösenden Aufgabe umfassend begründet.

[VE-R-RE-KR-02] Quantifizierung der Abdeckung des Anwendungsbereichs

Anforderung: Do

- Die Abdeckung des Anwendungsbereichs durch den Trainings-, Test- und Validierungsdatensatz wird, falls möglich, formalisiert und in einem quantitativen Maß festgehalten. Lässt sich der Eingaberaum beispielsweise als niedrig-dimensionaler Vektorraum und der Anwendungsbereich als Teilmenge davon formalisieren, so kann als einfachstes Maß überprüft werden, ob jede Zelle eines Grids in diesem Bereich Datenpunkte enthält.
  - Allgemein kann es sinnvoll sein, auf Verfahren zurückzugreifen, die zusätzliche Eingabedaten erzeugen, um beispielsweise eine bessere Abdeckung durch die angereicherten Trainingsdaten zu erreichen (vgl. **[VE-R-RO-MA-01]** und **[VE-R-RO-MA-02]**, sowie **[VE-R-RO-MA-04]** für Augmentierungstechniken). Häufig ist es dabei möglich, zusätzlich auf Kriterien zur Bewertung der »Neuheit« erzeugter Datenpunkte zurückzugreifen[66].
- Falls möglich, werden anwendungsspezifische Ziel-Intervalle für das zuvor gewählte Abdeckungsmaß des Anwendungsbereichs definiert und begründet.
- Sofern die angemessene Abdeckung des Anwendungsbereichs begründbar nicht quantitativ nachgewiesen werden kann, ist eine qualitative Argumentation zulässig. Insbesondere in Kontexten wie etwa Open-World-Anwendungen, in denen eine derartige Betrachtung nie abschließend sein kann, liegt eine Strategie zur Berücksichtigung (etwa Protokollierung) neuer Szenarien und Eingangsdaten während des Betriebs vor.

[VE-R-RE-KR-03] Qualität der Trainings- und Testdaten

Eine angemessene (quantitative) Abdeckung des Anwendungsbereichs durch Trainings- und Testdaten ist ein wesentlicher Faktor, um sicherzustellen, dass das ML-Modell alle möglichen Einsatzszenarien berücksichtigt und somit optimale Entscheidungsregeln lernt. Um eine zuverlässige Performanz der KI-Komponente im Anwendungsbereich sicherzustellen, sollten jedoch auch qualitative Anforderungen an die Daten formuliert werden, da diese einen nicht weniger relevanten Anteil an der Qualität des ML-Modells haben können. Wesentlich ist etwa der Wahrheitsgehalt der Daten bzw. die Korrektheit von Annotationen/Labels, damit die KI-Komponente aus den gelernten Zusammenhängen die richtigen Schlüsse zieht. Die Korrektheit der gespeicherten und verarbeiteten Daten ist insbesondere in Hinblick auf personenbezogene Daten erforderlich. Darüber hinaus können sich je nach Anwendungskontext weitere qualitative Anforderungen an die Daten ergeben, die beispielsweise im Zusammenhang mit technischen Einschränkungen stehen oder sich aus betrieblichen Prozessen und Anforderungen ergeben, wie etwa aus der Notwendigkeit des Reproduzierens von Ausgaben (vgl. **[TR-R-AF-KR-01]**) und damit einhergehender Eingaben, aus der Analyse von Fehler-/Unfallursachen (siehe z. B. **[SI-R-BD-MA-03]** und **[SI-R-FS-MA-14]**) oder aus der Auskunftsfähigkeit bezüglich Daten und Modellen (siehe auch **[TR-R-NB-MA-01]**, **[TR-R-EX-MA-01]** und **[DS-R-PD-MA-11]**).

---

66 Vgl. z. B. Odena, A. und I. Goodfellow (2018) TensorFuzz: Debugging Neural Networks with Coverage-Guided Fuzzing. Cornell University. https://arxiv.org/pdf/1807.10875.pdf (letzter Aufruf: 23.06.2021).

Anforderung: Do

- Es werden Kriterien zur Beurteilung der Datenqualität festgelegt. Bei der Wahl der Kriterien sollten mindestens die folgenden Aspekte berücksichtigt werden:
  - Technische Anforderungen (Format, Größe der Datei)
  - Vollständigkeit der Daten (z. B. Vorhandensein aller Attribute)
  - Wahrheitsgehalt der Daten
  - Korrektheit von Annotationen/Labels
  - Relevanz der Daten für den Anwendungsbereich
  - Verfügbarkeit/Zugriff auf Daten und Metadaten
- Für jedes der festgelegten Kriterien wird außerdem eine qualitative Zielsetzung/-eigenschaft formuliert, unter deren Einhaltung ein vertretbares Risiko bezüglich der Datenqualität hergestellt ist.
- Die Wahl der Kriterien und zugehörigen angestrebten Eigenschaften wird begründet. Ferner wird dargelegt, dass diese Wahl mit den Zielvorgaben in **[VE-R-RE-RI-01]** konform ist.

7.1.3 Maßnahmen

7.1.3.1 Daten

[VE-R-RE-MA-01] Ursprung und Qualität der Datenbasis

Anforderung: Do

- Die Herkunft der Trainings- und Testdaten wird dokumentiert und die Integrität der Datenquelle(n) eingeschätzt.
- Im Fall annotierter Daten liegt eine Dokumentation darüber vor, auf welche Weise die Annotationen bzw. Labels erstellt wurden. Außerdem wird beschrieben, wie sichergestellt ist, dass die Annotationen der Daten korrekt sind (z. B. durch Vier-Augen-Prüfung oder spezielle Software).
- Es wird dargelegt, dass die Daten qualitativ für das Training geeignet sind. Dabei wird insbesondere erläutert, inwiefern die Kriterien in **[VE-R-RE-KR-03]** erfüllt werden.
- Die »Kompatibilität« von Trainings- und Testdaten unter Beachtung des Anwendungsbereichs der KI-Komponente wird beschrieben.
  - Dabei ist insbesondere darauf einzugehen, ob die Daten strukturidentisch sind und der gleichen Verteilung angehören oder, ob es starke Abweichungen gibt.
  - Es wird beschrieben, durch welche Maßnahmen eine Überschneidung der Trainings- und Testdaten (*Data Leakage*) ausgeschlossen wird (z. B. durch lokal-sensitives *Hashing* der Daten oder statistische Methoden).

[VE-R-RE-MA-02] Wahl der Datenbasis

Anforderung: Do

- Die Wahl der Trainings- und Testdaten wird in Bezug auf die im Betrieb der KI-Anwendung zu erwartenden regulären Eingabedaten ausführlich begründet.
- Es wird dokumentiert, dass die Trainings- und Testdaten den Anwendungsbereich ausreichend abdecken. Dabei wird insbesondere auf folgende Punkte eingegangen:
  - Die Abdeckung des Anwendungsbereichs wird in der Dokumentation unter Angabe des in **[VE-R-RE-KR-02]** genannten Maßes und Zielintervalls (soweit möglich) quantitativ nachgewiesen. Für die Abdeckung der Anwendungsgrenze kann hierbei ggf. auf **[VE-R-RO-MA-01]** bis **[VE-R-RO-MA-03]** verwiesen werden.
  - Es liegt eine verständliche, aussagekräftige Dokumentation vor, in der ggf. ergriffene Maßnahmen zur besseren Abdeckung des Anwendungsbereichs, wie etwa Datenaugmentationen, beschrieben werden. Die Wahl dieser Maßnahmen wird begründet. Hierbei kann ggf. auf **[VE-R-RO-MA-01]**, **[VE-R-RO-MA-02]** oder **[VE-R-RO-MA-04]** verwiesen werden.
  - Sollte eine Quantifizierung der Abdeckung des Anwendungsbereichs nicht möglich sein, so muss eine ausführliche Begründung gemäß **[VE-R-RE-KR-02]** erfolgen.

- Für den Fall, dass die Trainingsdaten nicht den »realen« Daten im Betrieb entsprechen, weil sie z. B. durch ein anderes Verfahren erzeugt wurden, sollte die Evaluation der KI-Anwendung entsprechend angepasst sein. Die Evaluation sollte auf verwandten Datensätzen, die innerhalb des Anwendungsbereichs liegen, sich jedoch qualitativ von den Trainings- und Testdaten der Anwendung unterscheiden, durchgeführt werden (*Domain Adaptation Test Set*, z. B. Simulationsdaten, andere Aufnahmetechnik, andere Datenvorverarbeitung). Falls zutreffend, muss die Eignung der gewählten Evaluationsdatensätze begründet werden.

7.1.3.2 KI-Komponente

[VE-R-RE-MA-03] Wahl des Komponenten-Designs

Anforderung: Do

- Es liegt eine Dokumentation vor, die die Wahl der Modellkomponenten (Trainingsalgorithmus, Loss-Funktion, etc.) zu den gewählten Verlässlichkeitsanforderungen in Bezug setzt. Zudem wird begründet, warum Design und Architektur der KI-Komponente für den vorliegenden Anwendungsbereich angemessen sind. Falls verschiedene ML-Modelle für die KI-Anwendung in Betracht gezogen wurden, sind deren Abwägung sowie die Gründe für die getroffene Wahl zu beschreiben. Hierbei ist auch darauf hinzuweisen, falls zur Konfiguration des ML-Modells auf Frameworks zurückgegriffen wurde. Außerdem ist zu erläutern, wie die Features der Eingabedaten für das ML-Modell ausgewählt wurden.
- Im Fall, dass die Trainingsdaten nicht den »realen« Daten im Betrieb entsprechen, sind Maßnahmen zu beschreiben, die getroffen werden, um den Herausforderungen von *Concept-*, *Covariate-* und/oder *Prior Shift* zu begegnen (z. B. Methoden des *Transfer Learnings*) und die Generalisierbarkeit des Modells zu gewährleisten.

[VE-R-RE-MA-04] Systematische Schwachstellensuche

Anforderungen: Do | Te

- Es wird dokumentiert, inwiefern eine systematische Schwachstellensuche der KI-Komponente durch iterative Datenanpassung (*Closed Loop Testing*, siehe auch **[TR-R-EX-MA-06]**) oder introspektive Methoden (siehe **[TR-R-EX-MA-02]**) durchgeführt wurde. Falls dabei Schwachstellen identifiziert wurden, sind diese sowie die Maßnahmen, die daraufhin ergriffen wurden, festzuhalten.
  Beispiel: Eine Bildklassifikation, die Schiffe von anderen Entitäten trennt, wurde durch ein Heatmap-basiertes Verfahren analysiert. Die Analyse zeigte, dass für die Klasse »Schiff« das Wellenmuster auf dem Wasser ausschlaggebend war und nicht das Schiff als solches. Daraufhin wurde das ML-Modell auf einem augmentierten Datensatz neutrainiert, der über eine größere Vielfalt an Wellenmustern in den Bildern verfügt.

[VE-R-RE-MA-05] Verlässlichkeitstests der KI-Komponente

Anforderungen: Do | Pr | Te

- Es werden Tests der KI-Komponente auf im Training ungesehenen Daten (Testdaten) durchgeführt, die, wie in **[VE-R-RE-MA-02]** dargelegt, den Anwendungsbereich ausreichend abdecken. (Die Performanz der KI-Komponente unter besonderer Berücksichtigung der Grenze des Anwendungsbereichs wird im Risikogebiet **Robustheit** getestet, siehe **[VE-R-RO-MA-05]**.) Die Durchführung der Tests sowie die gemäß **[VE-R-RE-KR-01]** betrachteten Metriken und die damit erreichten Werte sind zu dokumentieren. Außerdem ist darzulegen, inwiefern die Tests gezielt nach Modellschwächen suchen.
  - Im Fall, dass Trainingsdaten aus einer anderen Domäne bzw. Verteilung verwendet wurden (etwa bei *Transfer Learning* auf synthetischen Daten), müssen die Testdaten dem eigentlichen Anwendungsbereich entsprechen (vgl. **[VE-R-RE-MA-01]** und **[VE-R-RE-MA-02]**).
- Sofern sich während der Entwicklung der KI-Anwendung Modellschwächen durch Verfehlen von Zielintervallen relevanter Metriken gemäß **[VE-R-RE-KR-01]** offenbart haben, sind diese ebenso wie die ergriffenen Korrekturmaßnahmen und die daraus gewonnenen Erkenntnisse zu dokumentieren.

7.1.3.3 Einbettung

[VE-R-RE-MA-06] Realtests der KI-Anwendung

Anforderungen: Do | Pr | Te

- Es werden umfangreiche Realtests der KI-Komponente (schon eingebettet und als KI-Anwendung funktionstüchtig) durchgeführt und dokumentiert. Dabei werden zum einen die gemäß **[VE-R-RE-KR-01]** relevanten Performanz-Metriken bestimmt, zum anderen einbettungsspezifische Anforderungen wie Laufzeitmaße und durch die Einbettung veränderte Eingabeverteilungen überprüft. (Die Performanz der KI-Anwendung unter besonderer Berücksichtigung der Grenze des Anwendungsbereichs wird im **Risikogebiet: Robustheit (RO)** getestet, siehe **[VE-R-RO-MA-06]**.)
  - Bei den Realtests ist insbesondere zu beachten, dass alle regulären Einsatzsituationen abgedeckt wurden. Falls möglich, kann dies durch Testen aller unter Realbedingungen möglichen Parameterkombinationen geschehen. Im Beispiel einer Anlage zum Sortieren von Äpfeln wären dies alle (kategorischen) Kombinationen von Farbe, Sorte und Größe. Sofern dargelegt werden kann, dass eine vollständige Abdeckung nicht möglich ist, wird eine sinnvolle Auswahl getroffen, die die wichtigsten Fälle exemplarisch abdeckt. Diese Auswahl ist im Kontext der Anwendung zu begründen.
- Modellschwächen, die bei Realtests aufgedeckt wurden, sind ebenso zu dokumentieren wie die ergriffenen Korrekturmaßnahmen und die daraus gewonnenen Erkenntnisse.

7.1.3.4 Maßnahmen für den Betrieb

[VE-R-RE-MA-07] Ergänzung zur Open-World-Abdeckung

Anforderungen: Do | Pr

- Falls die KI-Anwendung in einem Open-World-Kontext eingesetzt wird bzw. die vollständige Abdeckung des Anwendungsbereichs gemäß **[VE-R-RE-KR-02]** nicht gewährleistet werden kann, existiert ein dokumentierter Prozess zur Qualitätskontrolle neuer Eingangsdaten während des Regelbetriebs. Dieser ergänzt die regelmäßigen Überprüfungen aus **[VE-R-RO-MA-07]**. Außerdem ist dokumentiert, wie die Erkenntnisse aus diesem Prozess zu einer stetigen Anpassung und Verbesserung der KI-Anwendung führen (siehe beispielsweise **[VE-R-BD-MA-02]** sowie Ansätze des *Federated Learning* in **[DS-R-PD-MA-08]**). Sofern hierbei Daten protokolliert werden, ist die **Dimension: Datenschutz (DS)** zu berücksichtigen.

7.1.4 Gesamtbewertung

[VE-R-RE-BW] Gesamtbewertung

Anforderung: Do

- Unter Bezugnahme auf die ergriffenen Maßnahmen wird dargelegt, dass die in **[VE-R-RE-KR-01]** festgelegten Performanz-Metriken und die in **[VE-R-RE-KR-02]** festgelegten Überdeckungsmaße für den Anwendungsbereich in den jeweils dort definierten Zielintervallen liegen. Ferner wird begründet, dass die Anforderungen an die Datenqualität gemäß **[VE-R-RE-KR-03]** erfüllt sind.
- Sofern nicht alle in **[VE-R-RE-KR-01]** bis **[VE-R-RE-KR-03]** spezifizierten Anforderungen erfüllt werden, sind die Abweichungen zu dokumentieren. Dies gilt ebenfalls für nur teilerfüllte Anforderungen, bei denen etwa die Kriterien nicht, oder nicht immer, erreicht wurden.

7.2 Risikogebiet: Robustheit (RO)

Das Risikogebiet Robustheit trägt Risiken Rechnung, die durch eine geringe Änderung bzw. Störung eines Anwendungsfalls entstehen, den die KI-Anwendung unter normalen Umständen fehlerfrei behandeln soll. Diese Abweichungen können unterschiedlicher Natur sein, beispielsweise Bildverzerrungen, Sensorrauschen bzw. -ausfall oder unpräzise Datenerhebung wie Mess- oder Tippfehler. Eine besondere Fehlerklasse, die ebenfalls in dieses Risikogebiet fällt, sind adversariale Beispiele. Diese zeichnen sich durch eine geringe Abweichung von korrekt verarbeitbaren Eingangsdaten aus, die jedoch eine gravierende Abweichung vom erwarteten Ergebnis herbeiführen. Adversariale Beispiele können gezielt als Angriff entworfen sein (in diesem Fall werden sie als »Adversariale Attacke« bezeichnet), sind im Allgemeinen jedoch Ausdruck von Modellschwächen, die auch über Angriffsszenarien hinaus von Bedeutung sind.

Die hier betrachteten Abweichungen in den Eingabedaten bewegen sich allesamt in einem Rahmen, in dem eine korrekte Bearbeitung durch die KI-Komponente beabsichtigt ist. Dieser Rahmen bzw. diese Anwendungsgrenze müssen spezifiziert werden. Dazu werden zunächst mögliche Störungen erfasst und potenzielle Schäden abgeschätzt. Darauf aufbauend kann geprüft werden, ob das Gesamtrisiko der jeweiligen Störung angesichts des Einsatzkontextes vertretbar ist.

Ziel der für dieses Risikogebiet vorgesehenen Maßnahmen ist es, die Robustheit der KI-Komponente gegen die innerhalb des Anwendungsbereichs möglichen Abweichungen bzw. Störungen zu stärken und nachzuweisen, dass eine möglichst fehlerfreie Bearbeitung dieser abweichenden Eingaben erreicht wurde. Dadurch grenzt sich Robustheit insbesondere vom **Risikogebiet: Abfangen von Fehlern auf Modellebene (AF)** ab, das Störungen und Eingaben behandelt, bei deren Eintreten keine sinnvolle Bearbeitung durch die KI-Anwendung zu erwarten ist, oder diese nur unter unvertretbarem Risiko durchzuführen wäre. Im **Risikogebiet: Abfangen von Fehlern auf Modellebene (AF)** zielen die Maßnahmen entsprechend auf die Detektion dieser Eingaben ab anstatt auf die ordnungsgemäße Bearbeitung durch die KI-Komponente. In beiden genannten Risikogebieten sind die aufgeführten Maßnahmen ausschließlich auf die KI-Komponente bezogen. Darüber hinaus besteht die Möglichkeit, Störungen im Eingaberaum durch Maßnahmen abzufangen, die nicht die KI-Komponente, sondern stattdessen deren Einbettung betreffen (z. B. redundante Sensorauslegung oder Hardware-Monitoring). Diese finden sich in der **Dimension: Sicherheit (SI)** wieder.

7.2.1 Risikoanalyse und Zielvorgaben

[VE-R-RO-RI-01] Risikoabschätzung und Festlegen der Anwendungsgrenze

Anforderung: Do

Es liegt eine Dokumentation vor, die folgende Punkte qualitativ und konzeptionell adressiert:

- **Risikoabschätzung:** Es wird untersucht, welche Arten von Störungen angesichts des in **[VE-R-RE-RI-01]** spezifizierten Anwendungsbereichs zu erwarten sind.
  Beispiele möglicher Störungen sind:
  - Erwartbares Sensorrauschen, etwa bei einem optischen Sensor
  - Verzögerte Datenübertragung oder geringere Datenqualität (etwa für Echtzeitanwendungen)
    - Beispielsweise Audioübertragung bei einer KI-Anwendung zur Liveübersetzung
  - Verzerrung oder andersartige Verfremdungen von Eingabedaten
    - Typischerweise veränderte Objektlage bei der Objekterkennung, etwa falls ein Zielobjekt verrutscht oder nicht optimal positioniert ist
  - Erwartbare Änderung von Umgebungsverhältnissen
    - Wird die KI-Anwendung außerhalb kontrollierter Umgebungen genutzt, können sich externe Parameter etwa das Wetter ändern.
  - Adversariale Beispiele (etwa in Bild- oder Audiodaten)

  - Insbesondere relevant für KI-Anwendungen mit öffentlich zugänglicher Schnittstelle, dies umfasst sowohl eine öffentliche Online Nutzung (z. B. kostenfreies Serviceangebot) als auch eine Positionierung der Anwendung im öffentlichen Raum (z. B. Überwachungskamera)

  Die Liste ist eine Auswahl an Beispielen und erhebt keinen Anspruch auf Vollständigkeit. Es ist zu untersuchen, welche weiteren Störungen der KI-Anwendung im Anwendungsbereich auftreten können. Diese werden, falls möglich, formalisiert und in späteren Tests gezielt genutzt. Hierbei sei erneut auf das **Risikogebiet: Funktionale Sicherheit (FS)** verwiesen, welches insbesondere mögliche Fehlerquellen auf klassischer also KI-unabhängiger Seite betrachtet, die mithilfe konventioneller Algorithmik oder Maßnahmen behandelt werden können, beispielsweise defekte Pixel als Sensorfehler. Sofern diese Fehlerquellen auch mit KI-spezifischen Methoden behandelt werden, so sind diese und mögliche Überschneidungen mit der **Dimension: Sicherheit (SI)** ebenfalls hier zu dokumentieren.
  Es wird pro identifizierter Störungsart abgeschätzt, mit welcher Häufigkeit und in welcher Stärke (Intensität) diese potenziell auftritt. Darauf aufbauend wird für jede identifizierte Störung in Abhängigkeit der Stärke die Möglichkeit einer Fehleinschätzung bzw. eines Ausfalls der KI-Komponente analysiert und die daraus resultierenden Schäden abgeschätzt.
- **Festlegen der Anwendungsgrenze:** Basierend auf der Risikoabschätzung wird die Anwendungsgrenze spezifiziert. Die Anwendungsgrenze beschreibt den Übergang des Anwendungsbereichs in Eingabebereiche, für die keine sinnvolle Verarbeitung erwartet wird. Dieser Randbereich wird in Bezug auf zu erwartende Störungen festgelegt und beschreibt, bis zu welchem Störungsgrad eine Bearbeitung durch die KI-Komponente beabsichtigt ist. Eine Verarbeitung von (möglicherweise störungsbehafteten) Eingabedaten durch die KI-Komponente ist insbesondere nur dann zu beabsichtigen, wenn diese Eingabedaten einen sicheren Betrieb der KI-Anwendung erlauben. Deshalb sind bei Festlegung der Anwendungsgrenze auch die Risikoabschätzungen aus der **Dimension: Sicherheit (SI)** einzubeziehen. Anhand dieser Anwendungsgrenze soll sich zuordnen lassen, welche der zuvor gelisteten Störungen bis zu welcher Stärke im **Risikogebiet: Robustheit (RO)** (innerhalb der Anwendungsgrenze) und welche darüber hinaus im **Risikogebiet: Abfangen von Fehlern auf Modellebene (AF)** (außerhalb des Anwendungsbereichs) behandelt werden.
- **Zielvorgaben:** Für Störungen, die innerhalb der Anwendungsgrenze liegen, werden Zielvorgaben in Bezug auf die KI-Komponente gemacht, deren Erreichen bedeutet, dass das verbleibende Restrisiko akzeptabel ist. Die Zielvorgaben können sich nach Natur und Grad der Störung unterscheiden. Insbesondere sind sie mit den Anforderungen und Maßnahmen in der **Dimension: Sicherheit (SI)** abzustimmen, sodass die KI-Anwendung innerhalb der Anwendungsgrenze durch kombinierte Robustheits- und Sicherheitsmaßnahmen ausreichend abgesichert ist. Dies wird in der dimensionsübergreifenden Gesamtbewertung diskutiert.
  Beispiel: Ein Gesichtserkennungsmodell wird durch ein vorgehaltenes Foto getäuscht. Die KI-Anwendung könnte dies abfangen, indem sie Dynamiken wie etwa Blinzeln berücksichtigt. Eine Maßnahme aus der **Dimension: Sicherheit (SI)** hingegen wäre das zusätzliche Verwenden einer Infrarotkamera oder eines Lidar-Systems.

7.2.2 Kriterien zur Zielerreichung

Im Folgenden werden Methoden zur Quantifizierung der Robustheit und der Anwendungsgrenze gelistet, die objektiv überprüfbar machen, ob die gesetzten Ziele erreicht werden und die identifizierten Risiken akzeptabel sind.

[VE-R-RO-KR-01] Quantifizierung der Anwendungsgrenze

Anforderung: Do

- Die in **[VE-R-RO-RI-01]** beschriebene Anwendungsgrenze wird – soweit möglich – formalisiert, beispielsweise als Okklusionsgrad bei einer Fußgängererkennung oder als Signal-Rausch-Verhältnis bei einer *Conversational AI*. Da die Anwendungsgrenze ggf., wie in **[VE-R-RO-RI-01]** beschrieben, in Bezug auf verschiedene zu erwartende Störungen festgelegt wird, kann dabei eine quantitative Abstufung der verschiedenen Störungsgrade vorgenommen werden, z. B. falls für unterschiedliche Störungsgrade unterschiedliche Zielintervalle bezüglich der Performanz der KI-Anwendung vorgesehen sind.
- Mindestens sollten qualitative oder semantische Anforderungen an die Eingabedaten zur näheren Charakterisierung der Anwendungsgrenze festgelegt werden.

[VE-R-RO-KR-02] Quantifizierung der Robustheit

Anforderung: Do

- Die Robustheit der KI-Komponente bezüglich der Grenze des Anwendungsbereichs ist mit mathematisch-statistischen Maßen bzw. Metriken (vgl. **[VE-R-RE-KR-01]**) zu messen. Es ist zu dokumentieren und begründen, welche der Metriken im Folgenden zur Bewertung der Robustheit herangezogen werden.
- Für jede Störungsart innerhalb der Anwendungsgrenze werden Zielintervalle für die festgelegten mathematisch-statistischen Metriken definiert. Die Robustheit der KI-Komponente gegenüber diesen Störungen kann dann durch Einhalten des Zielintervalls in einem auf die jeweilige Störung abgestimmten Test objektiv überprüft werden. Die Zielintervalle können sich je nach Störungsrisiko und -grad unterscheiden, es muss jedoch begründet werden, warum die Wahl der Zielintervalle angemessen ist.

[VE-R-RO-KR-03] Abdeckung der Anwendungsgrenze

Anforderung: Do

- Die Abdeckung der Anwendungsgrenze wird, falls möglich, formalisiert und quantifiziert. Hierbei sollte von den Kriterien in **[VE-R-RE-KR-02]** ausgegangen werden, insbesondere, wenn Störungen der Datenpunkte durch Faktoren begünstigt oder beschrieben werden, die nicht umfänglich im **Risikogebiet: Verlässlichkeit im Regelfall (RE)** behandelt werden. Die Anwendungsgrenze kann zudem je nach Störungsart und -grad unterschiedlich abgestuft werden (vgl. **[VE-R-RO-KR-01]**). Lässt sich der Eingaberaum beispielsweise als niedrig-dimensionaler Vektorraum formalisieren, so kann als einfachstes Abdeckungsmaß überprüft werden, ob jede Zelle eines Grids im Bereich der Anwendungsgrenze Datenpunkte enthält.
- Es werden anwendungsspezifische Zielintervalle für die Abdeckung der Anwendungsgrenze definiert.

7.2.3 Maßnahmen

7.2.3.1 Daten

[VE-R-RO-MA-01] Daten für das Testen der Robustheit

Anforderung: Do

Es liegt eine Dokumentation vor, in der die Auswahl an Testdaten und deren Eigenschaften gemäß der folgenden Struktur beschrieben wird:

- Die Wahl der Datensätze zur Evaluation der KI-Komponente bezüglich möglicher Störungen wird begründet. Dabei ist darzulegen, wie durch die gewählten Datensätze die spezifischen Anforderungen des Anwendungsbereichs abgebildet werden und, dass die Daten quantitativ oder ggf. qualitativ bzw. semantisch für die in **[VE-R-RO-KR-01]** beschriebenen Störungen angemessen sind.
  - Die im Datensatz abgebildeten Störungen sollten in semantischer Nähe zu einem für die KI-Anwendung relevanten adversarialen Beispiel oder technischen Fehler liegen.

  - Sofern der Datensatz durch Augmentierung oder andere Verfahren, etwa zur Erzeugung adversarialer Attacken, gewonnen wird, ist zusätzlich zur Eignung des verwendeten Verfahrens (hinsichtlich Qualität, Nähe zur Störung etc.) auch die Eignung des (eventuell vorhandenen) zugrundeliegenden Datensatzes zu belegen. Hierzu kann beispielsweise auf **[VE-R-RE-MA-02]** zurückgegriffen werden.
  - Sofern die Datenquellen oder das verwendete Verfahren zur Datengewinnung auch in **[VE-R-RO-MA-02]** Verwendung fanden, ist zu erörtern, ob die Testdaten im Vergleich zu den zugehörigen Trainingsdaten eine hinreichende Störungsvariabilität aufweisen, damit ein mögliches *Overfitting* auf relevante Störungsmuster immer noch im Test entdeckt werden kann.
    Beispiel: Die Verwendung von etwa nur schwach variablen Augmentierungen (z. B. Bilddrehungen um 90° Winkel bei Bilderkennung) kann dazu führen, dass ein *Overfitting* auf die gewählte Augmentierung erfolgt (und somit keine Robustheit gegenüber beispielsweise Drehungen um 45° vorliegt).
- Es wird dokumentiert, wie die Integrität der Datenquelle(n) eingeschätzt wird. Falls dort bereits beschrieben, kann auf **[VE-R-RE-MA-01]** verwiesen werden.

[VE-R-RO-MA-02] Daten für robustes Training

Anforderung: Do

- Es liegt eine Dokumentation vor, aus der ersichtlich wird, ob spezifische Trainingsdaten verwendet werden, um eine erhöhte Robustheit zu erreichen (»robuste Trainingsdaten«). Hierzu können die in **[VE-R-RO-MA-01]** genannten Datenquellen herangezogen werden, sofern ihr Umfang eine sinnvolle Aufteilung in Trainings- und Testdaten zulässt. Alternativ können Augmentierungstechniken verwendet werden[67], um aus bestehenden Daten einen diverseren Datensatz zu gewinnen. Diese Techniken können auch bereits dynamisch in den Trainingsprozess eingebunden werden, siehe **[VE-R-RO-MA-04]**.
- Es wird untersucht, inwiefern die Wahl der Trainingsdaten für den Anwendungsbereich statistisch repräsentativ ist oder warum eine ggf. eingeschränkte Repräsentativität unkritisch oder gar angemessen ist. So könnte etwa ein Modell zur Kollisionserkennung gezielt mit einer im Vergleich zum Regel- und Realbetrieb überrepräsentativen Zahl von (Beinah-) Kollisionsbeispielen trainiert worden sein, um seine Performanz in seltenen, aber kritischen Situationen zu verbessern (siehe auch **[VE-R-RO-MA-03]**).

[VE-R-RO-MA-03] Untersuchung von *Corner Cases*

Anforderung: Do

- Es liegt eine Dokumentation vor, aus der eine Auseinandersetzung mit der Suche nach herausfordernden Eingabedaten (sogenannten *Corner Cases*) hervorgeht. Hierbei handelt es sich um »schwierige« Daten, die etwa auf Klassengrenzen liegen, oder Eingabedaten, die derart selten vorkommen, dass sie wahrscheinlich nicht in einem zufällig gewählten Datensatz zu finden sind, obwohl ihre korrekte Verarbeitung beabsichtigt ist.
  Beispiel: Für eine Bilderkennung im autonomen Fahren stellt ein Ball, der hinter einem parkenden Fahrzeug auf die Straße rollt, einen *Corner Cases* dar.

7.2.3.2 KI-Komponente

[VE-R-RO-MA-04] Entwicklungs- und Trainingsprozedur

Anforderung: Do

Es liegt eine Dokumentation vor, in der entlang der folgenden Struktur beschrieben wird, inwiefern die Entwicklung und das Training die Robustheit der KI-Komponente stärken:

---

67 Siehe beispielsweise: Hendrycks, D. et. al. (Dezember 2019). AugMix: A Simple Data Processing Method to Improve Robustness and Uncertainty. Cornell University https://arxiv.org/pdf/1912.02781.pdf (letzter Aufruf: 22.06.2021).

- Die möglichen in **[VE-R-RO-RI-01]** spezifizierten Störungen sind in Bezug zur Entwicklung bzw. zum Training des ML-Modells zu setzen. Dabei sind insbesondere die ergriffenen Maßnahmen, die zum angestrebten Grad an Robustheit beitragen, zu beschreiben, wie etwa
  - Augmentiertes Training (z. B. AugMix, adversariales Training[68]),
  - Regularisierung (z. B. *Label Smoothing*, *Self-Distilation*, *Dropout*, *Batch Normalization*),
  - Transfer Ansätze (z. B. vortrainierte *Backbones*, *Multi-Task Learning*),
- Es wird dargelegt, wie durch die ergriffenen Maßnahmen sowie die Wahl der Loss-Funktion(en) und des Trainingsalgorithmus
  - die Verlässlichkeit bzw. Robustheit gegenüber den genannten Störungen gefördert wird,
  - (falls zutreffend) der gewünschte Grad an Generalisierbarkeit erreicht wird, z. B. durch Ansätze zum *Multi-Task-* oder *Transfer-Learning*. Falls dort bereits beschrieben, kann hierbei auf **[VE-R-RE-MA-03]** verwiesen werden.

[VE-R-RO-MA-05] Testen der KI-Komponente hinsichtlich Robustheit

Anforderungen: Do | Pr | Te

- Unter Verwendung der in **[VE-R-RO-MA-01]** spezifizierten Datensätze werden Tests der KI-Komponente durchgeführt und dokumentiert, die die als relevant identifizierten Störungen abbilden. Die erreichten Zielwerte und die im Rahmen der Tests gewonnenen Erkenntnisse werden festgehalten.

7.2.3.3 Einbettung

[VE-R-RO-MA-06] Reale Generalisierungs-/Explorationstests der KI-Anwendung

Anforderungen: Do | Pr | Te

- In Ergänzung zu den im **Risikogebiet: Verlässlichkeit im Regelfall (RE)** durchgeführten Realtests wird die Robustheit der KI-Anwendung unter einbettungsspezifischen Anforderungen, wie veränderten Eingabeverteilungen und sonstigen absehbaren Störungen/Fehlern/Abweichungen, überprüft. Diese auf Störungen zielenden Generalisierungs-/Explorationstests werden durchgeführt und dokumentiert, während die KI-Komponente schon eingebettet und als KI-Anwendung funktionstüchtig ist. Insbesondere ist zu beschreiben, welche der in **[VE-R-RO-RI-01]** als relevant identifizierten Störungen durch diese Tests abgebildet werden.
- Es ist zu begründen, welche der zuvor festgelegten Zielintervalle (aus **[VE-R-RE-KR-01]** oder **[VE-R-RO-KR-02]**) zur Evaluierung der Generalisierungs-/Explorationstests herangezogen werden. Die erreichten Zielwerte und gewonnenen Erkenntnisse sind zu dokumentieren.

7.2.3.4 Maßnahmen für den Betrieb

[VE-R-RO-MA-07] Kontrolle der Eingabedaten im Betrieb

Anforderungen: Do | Pr | Te

- Nach Möglichkeit ist zu testen, ob die Eingabedaten Mindestanforderungen genügen (technische Qualität, korrektes Format) und zulässig sind (z. B. Bereinigung durch *Outlier Detection*). Abhängig von der Datenkomplexität ist auch die semantische Nähe zum Anwendungsfall zu überprüfen (für komplexe Eingaben wird eine mögliche Verletzung der Anwendungsgrenze in **[VE-R-AF-MA-08]** vertiefend betrachtet). Der Test zur Kontrolle der Eingabedaten wird im Produktivbetrieb kontinuierlich ausgeführt. Die Methoden, auf denen

---

68 Dazu siehe auch Carlini, N. et. al. (Februar 2019). On Evaluating Adversarial Robustness. Cornell University https://arxiv.org/pdf/1902.06705.pdf (letzter Aufruf: 22.06.2021); Kolter, Z. & Madry, A. (2021) Adversarial Robustness - Theory and Practice. https://adversarial-ml-tutorial.org/ (letzter Aufruf: 22.06.2021) und Zheng, S. et. al. (April 2016). Improving the Robustness of Deep Neural Networks via Stability Training. Cornell University. https://arxiv.org/pdf/1604.04326.pdf (letzter Aufruf: 22.06.2021).

dieser Test aufbaut, sowie mögliche Anschlussreaktionen, falls nicht bereits im **Risikogebiet: Funktionale Sicherheit (FS)** der Dimension Sicherheit beschrieben, werden dokumentiert und begründet.

[VE-R-RO-MA-08] Kontrolle der Ausgaben im Betrieb

Anforderungen: Do | Pr | Te

- Nach Möglichkeit ist ein Monitoring-Prozess bezüglich der Ausgaben der KI-Komponente im Produktivbetrieb zu etablieren (*Sanity Check*). Wird erkannt, dass Ausgaben die Anwendungsdomäne verlassen, sind diese abzufangen. Die technische Umsetzung des Monitoring-Prozesses ist zu dokumentieren und zu prüfen. **Beispiel:** Die Geschwindigkeit eines Fußgängers wird auf 35 km/h geschätzt. Diese zweifelhafte Prädiktion wird erkannt und führt zur Überprüfung durch eine redundante Einheit.

7.2.4 Gesamtbewertung

[VE-R-RO-BW] Gesamtbewertung

Anforderung: Do

- Unter Würdigung der in **[VE-R-RO-MA-05]** und **[VE-R-RO-MA-06]** durchgeführten Tests wird dargelegt, dass die KI-Komponente gemäß den Kriterien in **[VE-R-RO-KR-02]** robust ist. Ferner wird dokumentiert, dass die Kriterien **[VE-R-RO-KR-01]** und **[VE-R-RO-KR-03]** erfüllt sind.
- Sofern nicht alle in **[VE-R-RO-KR-01]** bis **[VE-R-RO-KR-03]** spezifizierten Anforderungen erfüllt werden, sind die Abweichungen zu dokumentieren. Dies gilt ebenfalls für nur teilerfüllte Anforderungen, bei denen etwa die Kriterien nicht, oder nicht immer, erreicht wurden.

7.3 Risikogebiet: Abfangen von Fehlern auf Modellebene (AF)

Während das **Risikogebiet: Robustheit (RO)** fehlerfreie Ausgaben der KI-Komponente bezüglich der Anwendungsgrenze zum Ziel hat, widmet sich das Risikogebiet Abfangen von Fehlern auf Modellebene Fällen, in denen ein Versagen der KI-Komponente absehbar oder unvermeidlich ist oder bei denen eine zu hohe Anzahl von Fehleinschätzungen zu unvertretbarem Risiko führt. Ein Beispiel für absehbares Versagen stellt ein optisches Erkennungssystem dar, das mit widrigen Aufnahmebedingungen konfrontiert ist. Ein unvertretbares Risiko kann beispielsweise durch einen nicht vorhergesehenen *Domain Shift* verursacht werden, etwa wenn eine Steuerungs-KI für eine autonome Drohne für den Einsatz auf offener Fläche in einer Häuserschlucht verwendet wird.

Der Übergang von Robustheit zu Versagen kann fließend sein. Ein gewisser Rauschpegel ist auch im normalen Betrieb nicht vermeidbar und fällt daher erst oberhalb eines Schwellwerts in diese Kategorie. Diese Anwendungsgrenze wurde bereits im **Risikogebiet: Robustheit (RO)** untersucht und so weit wie möglich spezifiziert.

Das Risikogebiet Abfangen von Fehlern auf Modellebene grenzt neben dem **Risikogebiet: Robustheit (RO)** auch an das **Risikogebiet: Funktionale Sicherheit (FS)**, da sich beide Kapitel mit der Überwachung des Funktionszustandes befassen. Dabei nähert sich das **Risikogebiet: Funktionale Sicherheit (FS)** von der Einbettungsseite, während in diesem Abschnitt von der KI-Komponente und insbesondere dem ML-Modell ausgegangen wird. Ansätze aus der **Dimension: Sicherheit (SI)** können hier, wo möglich, ergänzend referenziert werden. Eine Reihe von Einschränkungen, wie etwa oben genanntes Rauschen, kann mithilfe konventioneller Algorithmik (d. h. Maßnahmen aus dem **Risikogebiet: Funktionale Sicherheit (FS)**) erkannt werden. Oft jedoch ist das Anwendungsgebiet von KI zu breit, um es mithilfe klassischer Überwachung vollständig abzudecken. Beispielsweise ist anzunehmen, dass konventionelle Ansätze in der Regel nicht in der Lage sind, zu erkennen, ob ein*e Heimanwender*in versucht, ein Sprachassistenzsystem in einer anderen als der vorgesehenen Sprache zu bedienen. Um zu verhindern, dass die KI-Anwendung die unbekannte Eingabe fälschlicherweise als Anweisung zur Produktbestellung interpretiert, bedarf es eines Detektionsmechanismus auf Modellebene. Ein zentraler Fokus in diesem Risikogebiet ist daher die zuverlässige Erkennung problematischer Eingangsdaten, bei denen eine sinnvolle Verarbeitung nicht zu erwarten ist und deren ungehinderte Verarbeitung durch die KI-Komponente zu unvertretbarem Risiko in Hinblick auf einen Personen-, Sach- oder finanziellen Schaden führen kann. Ein Detektionsmechanismus auf Modellebene soll diese Eingaben abfangen und kann zudem Anschlussreaktionen auslösen.

Für die auf Ebene des ML-Modells zu implementierenden Detektionsstrategien kann nicht davon ausgegangen werden, dass sie zu anderen (oder auch nur ähnlichen) Problemstellungen generalisieren. Deshalb ist zunächst der Gültigkeitsbereich der Detektion abzugrenzen, was insbesondere bezüglich Störungen auf semantischer Ebene eine Herausforderung darstellt. Dabei lässt sich zwischen einem geschlossenen Anwendungsbereich und einem Open-World-Kontext unterscheiden. Ein Beispiel für Ersteres wäre eine statische Maschine am Förderband. Ein Beispiel für eine Open-World-Anwendung hingegen wäre eine *Smart Appliance*, die von nicht geschulten Nutzer*innen im Alltag verwendet wird, wie beispielsweise die zuvor genannte Spracherkennung. Im zweiten Fall sind die Daten naturgemäß breiter gestreut, aber auch ein geschlossener Anwendungsbereich schützt nicht vor unbekannten Eingangsdaten. Es könnte sich beispielsweise ein Mensch auf einem Förderband befinden.

7.3.1 Risikoanalyse und Zielvorgaben

[VE-R-AF-RI-01] Gültigkeitsbereich, Risikoanalyse und Zielvorgaben

Anforderung: Do

- **Gültigkeitsbereich:** Es wird dargelegt, welche Eingabedaten außerhalb des Anwendungsbereichs für das Abfangen von Fehlern auf Modellebene berücksichtigt werden sollen. Hierzu wird zunächst analysiert, inwiefern die ungehinderte Verarbeitung von *Out-of-Distribution*-Daten (OOD-Daten)[69] ein unvertretbares

69 Als *Out-of-Distribution*-Daten werden solche Eingabedaten bezeichnet, die außerhalb des Anwendungsbereichs liegen und nicht Teil der ursprünglichen Problemstellung sind.

Sicherheitsrisiko darstellt oder hohe finanzielle Schäden zur Folge hätte. Darauf aufbauend ist zum einen die in **[VE-R-RO-RI-01]** beschriebene Anwendungsgrenze heranzuziehen und festzulegen, welche ihrer Überschreitungen in diesem Risikogebiet, d. h. durch Detektionsmechanismen auf Modellebene, und welche (stattdessen oder zusätzlich) im **Risikogebiet: Funktionale Sicherheit (FS)** behandelt werden. Dieser Teil der Dokumentation erfolgt in Einklang mit der Risikoanalyse in **[SI-R-FS-RI-01]**. Zum anderen sollten auch Eingabedatenbereiche berücksichtigt werden, die weit jenseits der Anwendungsgrenze liegen, die aber dennoch auftreten könnten. Außerdem ist zu begründen, wenn für identifizierte OOD-Bereiche eine Detektion als nicht erforderlich erachtet wird.
Anmerkung: Die Betrachtung ist von der **Dimension: Sicherheit (SI)** (vgl. **[SI-R-FS-RI-01]**) insoweit abzugrenzen, als dort technisch bedingte Eingabevariabilität z. B. aufgrund von Sensorversagen behandelt werden kann. Der Fokus des Risikogebiets Abfangen von Fehlern auf Modellebene liegt auf inhärenten Eigenschaften der KI-Komponente, etwa der Fähigkeit, komplexe Eingabedaten zu verarbeiten. Problemstellungen wie etwa vollständiger Sensorausfall, die mithilfe konventioneller Verfahren behandelt werden können, sollten präferenziell in **Dimension: Sicherheit (SI)** adressiert werden.

- **Risikoanalyse:** Die Dokumentation ergänzt und erweitert die Risikoabschätzung der Eintrittswahrscheinlichkeiten und potenziellen Schäden in **[VE-R-RO-RI-01]** um die oben genannten Bereiche, für die eine Detektion auf Modellebene angestrebt wird. Der Schaden basiert hierbei auf der Annahme einer fehlerhaften Verarbeitung durch das Modell, sofern die Eingabe nicht abgefangen wird. Hierzu wird sowohl ein *Worst Case*, also die ungünstigste Ausgabe der KI-Anwendung, als auch zum Vergleich eine zufällige Ausgabe betrachtet. Am Beispiel einer Bildsegmentierung würde der erste Fall dem »Übersehen« kritischer Elemente entsprechen, etwa in der Computertomographie dem Fehlen einer Tumor-Segmentierung auf einem CT-Bild, der zweite Fall einer zufällig verzerrten Ausgabe, bei der alle Segmentierungsbereiche von ihrer üblichen Form abweichen und nur noch lose mit dem eigentlichen Bildinhalt korrespondieren.
- **Zielvorgaben:** Ausgehend von der Risikoanalyse werden qualitative Anforderungen bzw. Zielvorgaben an die Detektionsmaßnahmen gestellt, bei deren Erreichen das Risiko als beherrschbar eingestuft werden kann (unter Annahme geeigneter Anschlussreaktionen bei erfolgreicher Detektion). Diese Anforderungen werden im Folgenden genauer ausgeführt und geprüft.
  Die Wahl der Anschlussreaktionen an Detektionen auf Modellebene wird in der Risikoanalyse im **Risikogebiet: Funktionale Sicherheit (FS)** behandelt.

7.3.2 Kriterien zur Zielerreichung

Die starke Diversität der Problemstellungen in diesem Risikogebiet stellt eine Herausforderung dar, wenn es darum geht, Kriterien zu formulieren, die für beliebige KI-Anwendungen eine quantitative Beurteilung erlauben. Die Kriterien müssen daher an den jeweiligen Fall entsprechend den Zielvorgaben in **[VE-R-AF-RI-01]** angepasst werden. Dazu sollten mindestens die nachfolgenden Kriterien herangezogen werden.

[VE-R-AF-KR-01] *Out-of-Distribution*-Abdeckung

Anforderung: Do

- In Anlehnung an **[VE-R-RO-KR-03]** ist die Abdeckung der gemäß **[VE-R-AF-RI-01]** abzufangenden OOD-Bereiche des Eingaberaums formalisiert und quantifiziert. Falls eine Quantifizierung begründbar nicht möglich ist, kann, sofern dies ausreichend begründet wird, wie in **[VE-R-RO-KR-03]** auf eine qualitative kategorische Auseinandersetzung zurückgegriffen werden, etwa auf Basis von *Zwicky Boxes*.
- Insbesondere dann, wenn eine Formalisierung der Abdeckung begründbar auch qualitativ kategorisch nicht vollständig möglich ist oder aber in **[VE-R-AF-KR-03]** eine Generalisierungsfähigkeit gefordert wird, sollte die Abdeckung der OOD-Daten zusätzlich nach dem Vorhandensein von weiteren OOD-Datensätzen beurteilt werden. Bei diesen kann es sich etwa um »Rauschen«, bevorzugt aber um Daten für einen anderen verwandten Anwendungszweck handeln. Die Wahl des Datenbereichs ist zu begründen.

[VE-R-AF-KR-02] Vorhandensein von Mitigationsstrategien

Anforderung: Do

- Jedem zur Detektion vorgesehenen Eingabebereich aus **[VE-R-AF-RI-01]** ist, vergleichbar zu **[SI-R-FS-KR-03]**, eine Mitigationsstrategie gegenübergestellt, die ergriffen wird, sofern die Detektionsverfahren aus diesem Risikogebiet anschlagen. Dies kann beispielsweise die Übergabe der Kontrolle an den*die Nutzer*in sein. Eine Strategie muss nicht für alle genannten Bereiche anwendbar sein, sondern kann sich auch nur auf bestimmte Bereiche (oder sogar Teilbereiche) beziehen. Es ist jedoch jeder (Teil-)Bereich mit mindestens einer Maßnahme zu belegen. Diese Zuordnung ist tabellarisch festgehalten.

[VE-R-AF-KR-03] Anforderungen an die Detektionsmethoden

Anforderung: Do

- Für jeden der in **[VE-R-AF-RI-01]** genannten Bereiche werden, ausgehend von der Risikoanalyse sowie den gemäß **[VE-R-AF-KR-02]** zugeordneten Mitigationsstrategien, (falls möglich quantitative) Anforderungen an die Detektionsmethoden festgehalten. Hierbei wird mindestens auf die folgenden Punkte eingegangen:
  - Verlässlichkeit (Festlegen einer Metrik, mit der die Performanz der Detektionsmethoden gemessen werden kann, und eines Zielintervalls bzw. einer oberen Schranke, bis zu der ein Ausfall der Detektion vertretbar wäre. Dabei kann eine der Metriken aus **[VE-R-RE-KR-01]** gewählt werden.)
  - Reaktionszeit (Die maximal zulässige Detektionszeit wird abhängig von der anschließenden Mitigationsstrategie festgelegt.)
  - Einsatzschwelle (sofern der Übergang zwischen »robustem« Verhalten und Ausfall der KI-Komponente im Eingaberaum kontinuierlich ist, wird begründet ein Schwellwert für die Detektion festgelegt.)
  - Generalisierungsfähigkeit (Sofern die Abdeckung in **[VE-R-AF-KR-01]** qualitativ statt quantitativ ist, wird die Anforderung an die Generalisierungs- bzw. Extrapolationsfähigkeit der Detektionsmethode diskutiert. Generalisierungsfähigkeit bezeichnet in diesem Kontext die Eigenschaft der Detektionsmethode anzuschlagen, auch bei Eingabedaten außerhalb der Anwendungsgrenze, die nicht notwendigerweise im Betrieb zu erwarten sind und somit nicht explizit zur Konstruktion der Detektionsmethode einbezogen wurden.)
  - (Falls zutreffend) spezifische weitere Anforderungen, die sich aus der jeweiligen, in **[VE-R-AF-KR-02]** festgelegten anschließenden Mitigationsstrategie ergeben.

7.3.3 Maßnahmen

7.3.3.1 Daten

[VE-R-AF-MA-01] *Out-of-Distribution*-Datensatz

Anforderung: Do

- Es werden *Out-of-Distribution*-Datensätze (OOD-Datensätze) zum Testen der Detektionsmaßnahmen aufbereitet. Hierbei können ggf. aus dem vorhandenen Datenbestand in **[VE-R-RO-MA-01]** Daten aus relevanten OOD-Bereichen erzeugt werden. Diese sind derart mit weiteren Daten anzureichern, dass relevante Eingabedaten außerhalb des Anwendungsbereichs gemäß **[VE-R-AF-KR-01]** abgedeckt werden. Sofern die geforderte quantitativ oder qualitativ kategorische Abdeckung nicht erreicht werden kann, ist die Eignung der Testdatensätze zu begründen. Dies gilt in allen Fällen auch bezüglich seines Umfangs.

[VE-R-AF-MA-02] Datensatzsplits zur Extrapolation

Anforderungen: Do | Pr

- Sofern in **[VE-R-AF-KR-03]** eine Generalisierungsfähigkeit der Detektionsmethode gefordert wird oder die Abdeckung gemäß **[VE-R-AF-KR-01]** nicht erreicht werden kann, sind »künstliche« OOD-Daten auf Basis des Gesamtdatensatzes zu erzeugen. Zu diesem Zweck wird der Gesamtdatensatz so aufgeteilt, dass sich die entstehenden Subdatensätze strukturell voneinander unterscheiden. Für eine derartige Aufteilung können beispielsweise nachfolgende messbare Kriterien herangezogen werden:
  - semantische Eigenschaften, etwa auf Basis vorhandener Label-Informationen
  - statistische Eigenschaften, etwa Verteilungseigenschaften von Untermengen
  - latente Repräsentationen von Deep-Learning-Ansätzen, etwa von *Variational Auto Encodern* (bei der Verwendung von *DNN Features* sollten auch die Erfahrungen aus dem **Risikogebiet: Transparenz für Expert*innen (EX)** berücksichtigt werden)
- Die Wahl des Kriteriums für die Datensatz-Aufteilung ist zu begründen. Sie sollte sich in jedem Fall nach dem Anwendungskontext sowie der jeweils verwendeten Methode richten und dabei den Zweck der Gewinnung zusätzlicher »künstlicher« OOD-Datensätze berücksichtigen.
- Die so gewonnenen Datensatzsplits sind zu dokumentieren und können in **[VE-R-AF-MA-05]** für Extrapolationstests herangezogen werden.

7.3.3.2 KI-Komponente

[VE-R-AF-MA-03] Design zum korrelationsbasierten Abfangen von Fehlern in den Ausgaben

Anforderung: Do

- Auf der Ebene der KI-Komponente können bei komplexen Aufgabenstellungen zahlreiche Designansätze verwendet werden, die zu einer zuverlässigeren Erkennung unzulässiger Situationen führen. Eine Möglichkeit sind bestimmte Multi-Sensor-Ansätze, bei denen ein Ensemble aus ML-Modellen erstellt wird, die Informationen aus jeweils verschiedenen »Blickwinkeln« kombinieren. Ein anderer Design-Ansatz ist das *Multi-Label Learning*, bei dem ein ML-Modell trainiert wird, das verschiedene, aber zusammenhängende Aufgaben simultan löst, und dabei beispielsweise mehrere Labels ausgibt. Bei diesen Ansätzen gewinnt die KI-Komponente einerseits zusätzliche Robustheit, vgl. **[VE-R-RO-MA-04]**, Konflikte zwischen den Ausgaben können aber auch als Indikator für die Unzuverlässigkeit der Gesamtausgabe verwendet werden. Es ist darzulegen, inwiefern das Design der KI-Komponente das Abfangen von Fehlern auf Modellebene begünstigt, bzw. zu begründen, falls das Abfangen von Fehlern nicht explizit beim Design berücksichtigt wurde.

[VE-R-AF-MA-04] OOD-Tests

Anforderungen: Do | Pr | Te

- Anhand des OOD-Testdatensatzes aus **[VE-R-AF-MA-01]** wird statistisch untersucht, in welchem Anteil der Fälle die implementierten Detektionsmechanismen greifen. Diese Untersuchung kann als binäre Klassifikation (Erfolg/kein Erfolg) interpretiert werden und ist mindestens mit den zugehörigen Kriterien aus **[VE-R-AF-KR-03]** statistisch zu bewerten. Auch die dort genannten qualitativen Anforderungen an den Detektionsmechanismus sind zu überprüfen. Die Ergebnisse werden dokumentiert.
- Sollte während der Tests bei ggf. unzureichender Performanz eine (iterative) Anpassung des vorhandenen Datensatzes erfolgt sein bzw. zur Generierung oder Anforderung neuer Daten oder zur Anpassung der Maßnahme geführt haben, ist dies zu dokumentieren.

[VE-R-AF-MA-05] Extrapolationstest

Anforderungen: Do | Te

- Für den Fall, dass Methoden gemäß **[VE-R-AF-KR-03]** zu unbekannten/nicht hinreichend spezifizierbaren Daten generalisieren sollen, ist zu dokumentieren, dass die Tests aus **[VE-R-AF-MA-04]** auf den sich aus **[VE-R-AF-MA-02]** ergebenden Datensätzen durchgeführt wurden. Hierbei sollte jeweils ein Teil der Daten, getrennt nach den Kriterien des jeweiligen Splits, als außerhalb der Anwendungsdomäne betrachtet und dieselbe KI-Komponente erneut auf dem Rest der Daten trainiert worden sein. Ferner wird dokumentiert, inwiefern die Detektionsmethoden für die neu trainierten KI-Komponenten auf ihre Wirksamkeit im so entstandenen OOD-Bereich getestet wurden, und das Testergebnis festgehalten.
- Sofern nicht alle Splits verwendet wurden, weil ein Datensplit für eine der Detektionsmethoden als nicht relevant erachtet wird, etwa da das zu vermeidende Risiko prinzipiell nicht auftreten kann, wird dies zusammenfassend für die übergangenen Splits dokumentiert und begründet.
- Sollte während der Tests bei ggf. unzureichender Performanz der Detektionsmethoden eine (iterative) Anpassung des vorhandenen Datensatzes erfolgt sein bzw. zur Generierung oder Anforderung neuer Daten oder zur Anpassung der KI-Komponente bzw. ihrer Detektionsmethoden geführt haben, ist dies zu dokumentieren.

[VE-R-AF-MA-06] Unsicherheitsbewertung

Anforderung: Do

- Eine intrinsische Unsicherheitsbewertung der KI-Komponente kann als Form der Selbstbewertung prinzipiell dazu genutzt werden, Versagen aufgrund von Eingabedaten außerhalb des Anwendungsbereichs bzw. aufgrund von Unsicherheit des Modells zu detektieren. Sofern dies beabsichtigt ist, sind entsprechende Maßnahmen im **Risikogebiet: Einschätzung von Unsicherheit (UN)** zu ergreifen. In diesem Fall ist bei den Maßnahmen ein expliziter Bezug zum Risikogebiet Abfangen von Fehlern auf Modellebene herzustellen.
  - Bei Durchführung von zugehörigen Tests der Unsicherheitsbewertung im **Risikogebiet: Einschätzung von Unsicherheit (UN)** sollen insbesondere die hier spezifizierten OOD-Datensätze **[VE-R-AF-MA-01]** und Datensatzsplits **[VE-R-AF-MA-02]** berücksichtigt werden. Bei Auswertung der entsprechenden Ergebnisse und der abschließenden Bewertung in **[VE-R-UN-BW]** ist auf die Kriterien **[VE-R-AF-KR-03]** einzugehen.

7.3.3.3 Einbettung

[VE-R-AF-MA-07] Realtests

Anforderungen: Do | Te

- Soweit möglich wird die Detektion von Fehlermodi, die sich aus den gemäß **[VE-R-AF-RI-01]** abzufangenden OOD-Bereichen ergeben können, unter realen Bedingungen getestet. Dies ergänzt und erweitert die bestehenden Tests unter **[VE-R-RO-MA-06]** und findet unter gleichen Bedingungen und Anforderungen statt. Die Durchführung sowie die Ergebnisse des Realtests sind zu dokumentieren. Falls kein separater Test dieser Maßnahme vorliegt, ist darzulegen, dass dieser ggf. bereits durch **[SI-R-FS-MA-09]**, **[SI-R-FS-MA-11]** oder **[SI-R-FS-MA-13]** abgedeckt ist.

7.3.3.4 Maßnahmen für den Betrieb

[VE-R-AF-MA-08] Überwachung der Ein- und Ausgabedaten

Anforderungen: Do | Te

- Die bestehenden Maßnahmen zur Überwachung von Ein- und Ausgabedaten (siehe **[VE-R-RO-MA-07]** und **[VE-R-RO-MA-08]**) werden auf ihre Eignung zur Detektion von potenziellen Fehlerquellen auf Basis des OOD-Datensatzes aus **[VE-R-AF-MA-01]** geprüft. Sie können hierbei auf Modellebene durch weitere Verfahren der Vor- und Nachbereitung sowie des Monitorings ergänzt werden[70]. Die Ergebnisse und Anpassungen werden dokumentiert.

7.3.4 Gesamtbewertung

[VE-R-AF-BW] Gesamtbewertung

Anforderung: Do

- Unter Würdigung der in diesem Risikogebiet durchgeführten und dokumentierten Tests wird dargelegt, dass die Kriterien **[VE-R-AF-KR-01]** und **[VE-R-AF-KR-03]** erfüllt sind. Sofern Maßnahmen einander ergänzen, wird dargelegt, dass das Risiko eines korrelierten Versagens dieser Maßnahmen als beherrschbar angesehen werden kann.
- Ferner liegt eine tabellarische Übersicht vor, die den Detektionsmechanismen zum Abfangen von Fehlern aus diesem Risikogebiet entsprechende Mitigationsstrategien zuordnet. Ggf. kann hierzu auch auf **[SI-R-FS-MA-07]** verwiesen werden. Es wird dargelegt, dass diese Zuordnung im Einklang mit der Risikoanalyse und den Zielvorgaben in **[SI-R-FS-RI-01]** steht und die Anforderungen in **[VE-R-AF-KR-02]** erfüllt.
- Sollten die geplanten Detektionsmaßnahmen nicht realisierbar sein oder nicht ausreichen, um die Kriterien in diesem Risikogebiet zu erfüllen, so ist dies zu dokumentieren. Die hier nicht behandelbaren Problemstellungen können im **Risikogebiet: Funktionale Sicherheit (FS)** erneut betrachtet und das Restrisiko in der dimensionsübergreifenden Gesamtbewertung abgewogen werden.

---

70 Im Fall von Segmentierungs-Tasks könnte beispielsweise MetaSeg in Betracht gezogen werden, siehe: Rottmann, M. et. al. (November 2018). Prediction Error Meta Classification in Semantic Segmentation: Detection via Aggregated Dispersion Measures of Softmax Probabilities. Cornell University. https://arxiv.org/pdf/1811.00648.pdf (letzter Aufruf: 22.06.2021).

7.4 Risikogebiet: Einschätzung von Unsicherheit (UN)

Das Risikogebiet Einschätzung von Unsicherheit soll sicherstellen, dass die KI-Anwendung, falls erforderlich, eine zutreffende Aussage über die Konfidenz ihrer Ausgaben trifft. Eine realistische Unsicherheitsbewertung durch die KI-Komponente erlaubt es, die Risiken bei Weiterverarbeitung der prädizierten Ausgaben vorherzusehen und anwendungsfallspezifisch zu reagieren. Ähnlich wie die eigentlichen Ausgaben einer KI-Komponente sind die Konfidenzen mit Fehleinschätzungen belegt, sodass das hierdurch entstehende Risiko abgeschätzt werden muss.

Die Selbsteinschätzung von Unsicherheit durch die KI-Komponente ist nicht für alle Anwendungsbereiche erforderlich. Für komplexere Anwendungen, insbesondere im Open-World-Kontext, kann sie jedoch hilfreich oder sogar erforderlich sein. Selbst fehlerbehaftete Unsicherheitsschätzungen können in manchen Fällen durch Kalibrierung nachträglich korrigiert und auf realistische Größen abgebildet werden. So kann eine (korrekt) kalibrierte Unsicherheitsschätzung dazu beitragen, unübliche Situationen frühzeitig zu erkennen oder auch ein Verlassen der Anwendungsdomäne anzuzeigen. Während eine Reihe an Verfahren bereits intrinsisch Unsicherheitsmaße beinhalten (z. B. die probabilistische Ausgabe einer DNN-basierten Bildklassifikation), sind diese meist jedoch schlecht kalibriert. Gerade im Fall Neuronaler Netze ist es die Regel, dass intrinsische Unsicherheitsmaße Risiken deutlich unterschätzen und hohe Konfidenzen auch bei offensichtlicher Fehlklassifikation vorgeben. Im Rahmen dieses Risikogebiets soll daher zunächst evaluiert werden, inwieweit für die gegebene KI-Anwendung ein Unsicherheitsmaß erforderlich ist und in welcher Güte es vorliegen muss.

Beispiel: Die Güte des Endprodukts einer automatisierten Fertigungsstraße kann entweder durch aufwendige Qualitätskontrolle durch Experten oder durch eine KI-Anwendung unter Nutzung geeigneter Sensordaten bestimmt werden. Während ersteres Verfahren kostspielig ist, liefert Letzteres weniger verlässliche Schätzungen. Eine Unsicherheitsbewertung kann den Nutzen der KI-Anwendung zur Qualitätskontrolle steigern, indem es diejenigen Produkte markiert, bei denen eine zusätzliche manuelle Prüfung lohnend sein könnte. Das Niveau der Unsicherheitsschätzung sollte hinreichend gut sein, sodass weder zu viel Ausschuss entsteht noch ein Übermaß an Tests erforderlich wird.

7.4.1 Risikoanalyse und Zielvorgaben

[VE-R-UN-RI-01] Festlegung und Darstellung eines Unsicherheitsmaßes

Anforderung: Do

- **Risikoanalyse:** Es wird untersucht, inwiefern das sich aus den vorangegangenen Risikogebieten ergebende Restrisiko bezüglich Verlässlichkeit durch realistische Unsicherheitsschätzung weiter mitigiert werden kann. Dabei ist auf den spezifischen Anwendungsfall einzugehen. Wenn Unsicherheit gemäß **[VE-R-AF-MA-06]** zum Abfangen von Fehlern auf Modellebene genutzt wird, sind bei dieser Analyse auch die Anschlussreaktionen zu berücksichtigen, die bei signalisierter hoher Unsicherheit einsetzen sollen. Falls ein Unsicherheitsmaß im vorliegenden Anwendungskontext als nicht erforderlich oder möglich erachtet wird, ist dies ausführlich zu begründen.
- **Zielvorgaben:** Ausgehend von der Risikoanalyse wird festgelegt, ob eine Unsicherheitsschätzung in die KI-Anwendung implementiert werden soll. Ist dies der Fall, so werden Zielvorgaben formeller sowie qualitativer Natur gemacht, die sicherstellen, dass die geplante Unsicherheitsschätzung zur Mitigation bestehender Restrisiken hinsichtlich der Verlässlichkeit beiträgt.
  - Die formellen Zielvorgaben sollen insbesondere die Ausgabe der Unsicherheitsschätzung spezifizieren, z. B. ob kategorisch oder probabilistisch. Unter Ersterem ist eine Einstufung als diskrete Level (sicher/unsicher oder niedrig/mittel/hoch) zu verstehen, während der probabilistische Ansatz möglichst akkurate Wahrscheinlichkeitsaussagen über das Versagen der KI-Komponente bei gegebenem Input machen soll.

7.4.2 Kriterien zur Zielerreichung

Aus der Risikoanalyse folgen eine oder mehrere primäre Anforderungen, die die Unsicherheitsschätzung erfüllen muss. Diese betreffen ihre Güte, gewissermaßen die Zuverlässigkeit der Unsicherheitsschätzung selbst, messbar durch eine geeignete (Kalibrierungs-)Metrik. Hierbei ist dem Umstand Rechnung zu tragen, dass Unsicherheitsschätzungen einem Bias unterliegen können, sodass sich falsche Unsicherheitsschätzungen in bestimmten Anwendungsfällen häufen. Beispielsweise könnte eine Personenerkennung die Konfidenz bei der Erkennung von hell gekleideten Personen überschätzen, im Durchschnitt aber dennoch die Qualitätsanforderungen an die Unsicherheitsschätzung erfüllen. Wege, diese Missstände zu messen und die hieraus resultierenden Risiken zu minimieren, müssen betrachtet werden.

[VE-R-UN-KR-01] Dokumentation des Unsicherheitsmaßes und dessen Güte

Anforderung: Do

- Es werden mindestens eine Metrik zur Bewertung der Qualität der Unsicherheitsschätzung sowie Zielintervalle festgelegt. Die Wahl soll mit den in **[VE-R-UN-RI-01]** beschriebenen Zielvorgaben korrespondieren und ist zu begründen. Als Metrik sollte eine der unten aufgelisteten Größen gewählt werden. Sollte eine davon abweichende Metrik verwendet werden, ist diese ausführlich zu beschreiben und die abweichende Wahl zu begründen.
- Zusätzlich können semantische Dimensionen des Anwendungsfalls festgelegt werden, entlang derer die Qualität der Unsicherheitsschätzung konstant sein sollte. Falls etwa bekannt ist, dass die Performanz des Modells für einen bestimmten Teilbereich der Daten gering ist, so sollte sich dies auch in der Unsicherheitsschätzung widerspiegeln.
  Beispiel: Wenn eine Gesichtserkennung auf Bildern von People of Color eine erhöhte Fehlerrate aufweist, dann sollte die Unsicherheitsschätzung dazu korrelierend erhöht sein. Diese semantischen Dimensionen sind bei nachfolgenden Tests speziell zu untersuchen, um zu vermeiden, dass sich etwa Unsicherheitsfehlschätzungen für bestimmte Fälle häufen.

In der Literatur diskutierte Arten der Bewertung von Unsicherheitsmaßen sind u. a.:

- Die *Negative Log-Likelihood* bewertet eine probabilistische Ausgabe der KI-Komponente und akkumuliert für jedes Beispiel des Testdatensatzes die »Wahrscheinlichkeit«, das zugehörige korrekte Label unter der prädizierten Wahrscheinlichkeitsverteilung zu ziehen.
- Für Klassifikationen ist der *Brier-Score* verbreitet, der den quadrierten euklidischen Abstand der Konfidenzausgabe auf dem Simplex (im Falle binärer Klassifikation also dem Intervall [0,1]) zum als korrekt angesehenen Label als Ecke des Simplex (binär durch die Menge {0,1} gegeben) berechnet und für den Testdatensatz akkumuliert.
- Der *Expected Calibration Error* (ECE) stellt dar, ob Aussagen mit z. B. 90 Prozent Konfidenz auch in 90 Prozent der Fälle korrekt sind. Dieses Verfahren könnte auch auf kategorielle Unsicherheits-Scores angepasst werden, sofern diese zu Wahrscheinlichkeitsintervallen korrespondieren.
- Für kategorische Unsicherheitsangaben muss i. d. R. auf ein im Detail zu dokumentierendes, meist heuristisches Bewertungsschema zurückgegriffen werden, das dem entstehenden Risiko Rechnung trägt.

7.4.3 Maßnahmen

7.4.3.1 Daten

[VE-R-UN-MA-01] Wahl eines mit Unsicherheiten annotierten Datensatzes

Anforderung: Do

- In einigen Fällen kann Unsicherheit bereits eine Eigenschaft des *Ground-Truth*-Datensatzes sein. Beispielsweise könnten für die Segmentierung eines medizinischen Bilddatensatzes mehrere Expert*innen (leicht) abweichende Labels bereitstellen oder es könnte aufgrund der Aggregation komplexer bzw. großer

Datenmengen intrinsische (sog. aleatorische) Unsicherheit vorliegen, etwa in Bezug auf gemittelte Datenpunkte und ihre zugehörige Standardabweichung. Steht ein Datensatz zur Verfügung, der Informationen über die Unsicherheit von Labels enthält, so ist zu dokumentieren, ob und auf welche Art bzw. in welchem Umfang die KI-Komponente mit diesem Datensatz trainiert wurde, und in welcher Form die Unsicherheitsannotation verwendet wurde. Weiterhin ist der Prozess zur Erhebung oder Schätzung der Unsicherheitsannotationen zu dokumentieren und eine Abschätzung der Güte dieser Annotationen anzugeben.
Anmerkung: Datensätze dieser Form sind nur in seltenen Fällen verfügbar. Diese Maßnahme ist daher nur beschränkt anwendbar, kann jedoch einen deutlichen Mehrwert liefern.

7.4.3.2 KI-Komponente

[VE-R-UN-MA-02] Wahl einer geeigneten Methode zur Unsicherheitsbewertung

Anforderung: Do

- Abhängig von dem verwendeten ML-Modell können unterschiedliche Ansätze mit unterschiedlicher Güte und ggf. unterschiedlichem Aufwand zur Implementierung einer Unsicherheitsschätzung gewählt werden. Ausgehend von den Zielvorgaben und Kriterien ist die entsprechende Wahl zu begründen, wobei sowohl positive als auch negative Eigenschaften der Methode darzulegen sind. Etablierte Verfahren zur Unsicherheitsbewertung sind beispielsweise (Kombinationen von):
  - Bayesschen Netzen (z. B. *Monte-Carlo Dropout*[71]),
  - *Parametric Uncertainty*[72],
  - *Deep Ensembles*[73].

[VE-R-UN-MA-03] Post-Processing zur Kalibrierung

Anforderung: Do

- Zur Verbesserung der Unsicherheitsschätzung können Verfahren implementiert werden, die die Konfidenzausgabe zum Zweck einer besseren Kalibrierung nachverarbeiten. Die Wahl des Post-Processing-Verfahrens ist ausgehend von den Qualitätskriterien zu begründen und zu dokumentieren. Etablierte Verfahren zur Nachverarbeitung von Konfidenzausgaben sind beispielsweise *Temperature Scaling*[74] oder *Isotonic Regression*[75].

---

71 Gal, Y. & Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In international conference on machine learning. pp. 1050–1059.

72 Nix, D. A. and Weigend, A. S. (1994). Estimating the mean and variance of the target probability distribution. In Proceedings of IEEE International Conference on Neural Networks 1994, volume 1, pp. 55–60. IEEE

73 Lakshminarayanan, B., Pritzel, A. & Blundell, C., (2017). Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In I. Guyon and U. V. Luxburg and S. Bengio and H. Wallach and R. Fergus and S. Vishwanathan and R. Garnett, ed. Advances in Neural Information Processing Systems. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/9ef2ed4b7fd2c810847ffa5fa85bce38-Paper.pdf. (letzter Aufruf: 22.06.2021).

74 Guo, C. et al., 2017. On Calibration of Modern Neural Networks. In Precup, Doina and Teh, Yee Whye, ed. Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine Learning Research. PMLR, pp. 1321–1330. http://proceedings.mlr.press/v70/guo17a.html (letzter Aufruf: 22.06.2021).

75 Niculescu-Mizil, A. & Caruana, R. (2005). Predicting good probabilities with supervised learning. In Proceedings of the 22nd international conference on Machine learning (ICML '05). Association for Computing Machinery, New York, NY, USA, 625–632. https://doi.org/10.1145/1102351.1102430 (letzter Aufruf: 22.06.2021).

[VE-R-UN-MA-04] Testen der Unsicherheitsschätzung

Anforderungen: Do | Pr | Te

- Die Unsicherheitsschätzung wurde auf Daten getestet, die weder im Training noch zur Kalibrierung verwendet wurden. Falls in **[VE-R-UN-KR-01]** bestimmte semantische Dimensionen vorgegeben wurden, sind die Testdaten so zu wählen, dass die Unsicherheitsschätzung entlang dieser Dimensionen untersucht werden kann. Ferner sollten die Anforderungen in **[VE-R-AF-MA-06]** berücksichtigt werden, falls Unsicherheitsschätzung als Detektionsmethode zum Abfangen von Fehlern verwendet wird. Zur Bewertung der Qualität werden die in **[VE-R-UN-KR-01] f**estgelegten Metriken und Zielintervalle herangezogen. Die Testergebnisse sind zu dokumentieren.
- Sollte während der Tests bei ggf. unzureichender Qualität eine (iterative) Anpassung der Unsicherheitsschätzung erfolgt sein (z. B. durch zusätzliche Post-Processing-Maßnahmen), ist dies zu dokumentieren.

7.4.3.3 Einbettung

[VE-R-UN-MA-05] Überprüfung der Anschlussreaktionen

Anforderungen: Do | Te

- Für den Fall, dass die Unsicherheitsschätzung Anschlussreaktionen auslösen kann bzw. soll, sind diese und deren Implementierung zu dokumentieren, ggf. in Abstimmung mit dem **Risikogebiet: Abfangen von Fehlern auf Modellebene (AF)** (vgl. **[VE-R-AF-MA-06]** und **[VE-R-AF-KR-02]**). Außerdem werden Realtests durchgeführt, in denen das Auslösen der Anschlussreaktionen durch die Unsicherheitsschätzung gezielt provoziert und getestet wird. Sofern die in dieser Maßnahme geforderten Tests bereits an anderer Stelle dokumentiert sind, beispielsweise in **[VE-R-AF-MA-07]** oder **[SI-R-FS-MA-13]**, kann stattdessen auf diese Stelle verwiesen werden.

7.4.3.4 Maßnahmen für den Betrieb

Für diese Kategorie sind keine Maßnahmen vorgesehen.

7.4.4 Gesamtbewertung

[VE-R-UN-BW] Gesamtbewertung

Anforderung: Do

- Es liegt eine Dokumentation darüber vor, dass die in **[VE-R-UN-KR-01]** aufgeführten Kriterien erreicht wurden.
- Sofern nicht alle in **[VE-R-UN-KR-01]** spezifizierten Anforderungen erfüllt werden, sind die Abweichungen zu dokumentieren. Dies gilt ebenfalls für nur teilerfüllte Anforderungen, bei denen etwa die Kriterien nicht, oder nicht immer, erreicht wurden.

7.5 Risikogebiet: Beherrschung der Dynamik (BD)

Das Risikogebiet Beherrschung der Dynamik behandelt Risiken des *Model-* und *Concept Drifts*. Dadurch wird sichergestellt, dass die Verlässlichkeit der KI-Anwendung auch während des Betriebs aufrechterhalten wird.

Zum einen kann die Verlässlichkeit der KI-Anwendung dadurch geschwächt werden, dass sich äußere Umstände ändern. Falls sich beispielsweise die statistischen Eigenschaften der zu prädizierenden Größe verändern, könnte das ML-Modell nicht mehr geeignet sein, um diese Größe optimal zu beschreiben, und würde dementsprechend an Performanz einbüßen. Dies könnte z. B. bei einer KI-Anwendung der Fall sein, die vor Beginn der Covid-19-Pandemie auf die Erkennung von Gesichtern trainiert wurde und nun mit Gesichtern konfrontiert ist, die zum Teil mit Masken bedeckt sind. Aber auch Veränderungen der Rahmenbedingungen wie beispielsweise Gesetzesänderungen können nach Inbetriebnahme der KI-Anwendung Maßnahmen erfordern.

Um die Verlässlichkeit der KI-Anwendung zu jedem Zeitpunkt im Produktivbetrieb zu gewährleisten, sollte die korrekte Funktionsweise regelmäßig und in angemessenen Abständen überprüft werden. Ferner können geeignete Maßnahmen etabliert werden, etwa das Abspeichern herausfordernder Szenarien im Produktiveinsatz, um die Verlässlichkeit der KI-Anwendung stetig zu erhöhen.

Zusammenfassend ergeben sich in diesem Risikogebiet die folgenden zwei Risikokategorien:

1. ***Model Drift*:** Die KI-Anwendung hat nach erneutem Training auf während des Betriebs neu aufgenommenen Trainingsdaten eine verringerte Verlässlichkeit.
2. ***Concept Drift*:** Veränderte äußere Bedingungen stellen neue Anforderungen an die Verlässlichkeit.

7.5.1 Risikoanalyse und Zielvorgaben

[VE-R-BD-RI-01] Risikoanalyse und Zielvorgaben

Anforderung: Do

- **Risikoanalyse:** Es ist dokumentiert, ob und in welchem Umfang die KI-Anwendung während des Betriebs weiterlernt. Ist dies der Fall, wird ferner dargelegt, welche Anforderungen an die neu einkommenden Trainingsdaten gestellt werden und welche Prozesse oder Mechanismen zur Kontrolle der neu einkommenden Daten bestehen. Darauf aufbauend wird die Wahrscheinlichkeit falschen Weiterlernens im Betrieb und daraus potenziell resultierender Schäden abgeschätzt. Ferner wird analysiert, welche Arten von *Concept Drift* absehbar oder potenziell auftreten können, und es wird dokumentiert, welche Konsequenzen oder Schäden sich möglicherweise aus der Nichteinhaltung externer (möglicherweise veränderter) Anforderungen ergeben können.
- **Zielvorgaben:** Es werden Ziele für die Erkennung, Erfassung und Behandlung neu entstehender Fehlerfälle im Betrieb festgelegt, bei deren Einhaltung ein vertretbares Risiko hergestellt ist.

7.5.2 Kriterien zur Zielerreichung

[VE-R-BD-KR-01] Intervalle und Qualitätsanforderungen für Prüfung im Betrieb

Anforderung: Do

- Es werden angemessene Prüfintervalle für die KI-Anwendung zur Beurteilung der Verlässlichkeit gemäß der in den vorangegangenen Risikogebieten der Dimension Verlässlichkeit gewählten Metriken und Zielintervalle festgelegt und dokumentiert. Die Priorisierung der Risikogebiete in den geplanten regelmäßigen Tests ist anwendungsspezifisch vorzunehmen und ausführlich zu begründen. Die Länge der Prüfintervalle ist insbesondere mit der erwarteten Geschwindigkeit von relevanten *Concept Drifts* in Beziehung zu setzen (z. B. könnte bei einer Erkennung von Verkehrsteilnehmer*innen entschieden werden, einmal im Jahr zu überprüfen, ob neue Verkehrsteilnehmer*innen, wie bspw. Personen auf E-Scootern, ausreichend genau detektiert werden).

- Es werden (quantitative) Kriterien (z. B. Schwellwerte bezüglich der Performanz) festgelegt und dokumentiert, die beschreiben, in welchen Szenarien eine Neuevaluierung der KI-Anwendung, ggf. verbunden mit erneutem Training, erforderlich ist.
- Ferner werden Kriterien erarbeitet und dokumentiert, um die Qualität von Prozessen zur Erfassung neuer Anwendungs- und Fehlerfälle sowie deren Behandlung zu beurteilen.
- Es wird dargelegt, dass die gewählten Kriterien im Einklang mit den Zielvorgaben **[VE-R-BD-RI-01]** sind.

7.5.3 Maßnahmen

7.5.3.1 Daten

Für diese Kategorie sind keine Maßnahmen vorgesehen.

7.5.3.2 KI-Komponente

Für diese Kategorie sind keine Maßnahmen vorgesehen.

7.5.3.3 Einbettung

Für diese Kategorie sind keine Maßnahmen vorgesehen.

7.5.3.4 Maßnahmen für den Betrieb

[VE-R-BD-MA-01] Vermeidung von *Catastrophic Forgetting* auf neuen Trainingsdaten

Anforderung: Pr

- Es ist ein Prozess etabliert, der bei Verwendung neuer Trainingsdaten, z. B. bei inkrementellem Training einer KI-Komponente über mehrere Zyklen hinweg, überprüft, dass die inkrementellen Trainingsschritte nicht zu Performanz-Verlusten auf der vorherigen Datenbasis führen, sofern diese für die Anwendung noch relevant ist.
- Die neuen Trainingsdaten und Modellversionen sollten nach jedem inkrementellem Trainingsschritt gespeichert werden. Ferner sollte die Verteilung der neuen Trainingsdaten analysiert werden und die KI-Anwendung sowohl auf den neuen als auch auf alten Daten getestet werden. Weitere Maßnahmen wie etwa eine lineare Kombination von Ausgaben der verschiedenen Modellversionen sind denkbar, um dem Vergessen zuvor erlernter Muster/Modelle entgegenzuwirken.
- Die genauen Abläufe des etablierten Prozesses werden dokumentiert.

[VE-R-BD-MA-02] Neulernen bei Vorliegen neuer Trainingsdaten

Anforderung: Pr

- Es ist ein Prozess etabliert, der bei Vorliegen neuer Trainingsdaten (z. B. wegen der Notwendigkeit der ständigen Aktualität oder aufgrund von neuen Daten im Zusammenhang eines *Concept Drifts*) ausgehend von **[VE-R-BD-KR-01]** ein systematisches Neutraining/Nachlernen des Modells unter Einhaltung aller Trainings- und Testanforderungen dieses Katalogs veranlasst.
- Die Notwendigkeit neuer Kategorien oder Datenerhebungsverfahren wird für jede neue Datenerhebung geprüft, dokumentiert und implementiert.
- Die genauen Abläufe des etablierten Prozesses werden dokumentiert.

[VE-R-BD-MA-03] Regelmäßige Überprüfung der KI-Anwendung

Anforderungen: Do | Pr

- Es gibt einen Prozess, der eine gemäß den in **[VE-R-BD-KR-01]** definierten Prüfintervallen periodische Modellüberwachung hinsichtlich Verlässlichkeit gewährleistet. Der Prozess kann eine Kombination aus menschlicher Überprüfung in regelmäßigen Abständen, z. B. durch den*die Nutzer*in oder Mitarbeitende aus der IT-Abteilung, und kontinuierlichem automatischen Monitoring sein.
  - Im Rahmen dieser Überwachung, welche ggf. durch **[VE-R-AF-MA-08]** ergänzt wird, erfolgt auch eine Überprüfung, ob sich die Verteilung der Eingabedaten im Betrieb verändert. Zur Detektion von Veränderung, etwa der Eingabedaten oder einer Performanz-Metrik, kann – abhängig von der Komplexität der Variable – ein (*online*) *Drift-Detection*-Algorithmus implementiert werden, wie beispielsweise ADWIN (*Adaptive Windowing*)[76]. Die Wahl des Verfahrens ist zu begründen.
  - Kritische und neuartige Input-Daten, die beispielsweise *Concept Drifts* wiedergeben, werden dauerhaft abgespeichert mit dem Ziel, die Verlässlichkeit etwa durch weiteres Training oder Verbesserung der Detektionsmaßnahmen (siehe **Risikogebiet: Abfangen von Fehlern auf Modellebene (AF)**) zukünftig zu erhöhen. Die Speicherung erfolgt in Übereinstimmung mit der **Dimension: Datenschutz (DS)**.
  - Werden die Verlässlichkeitsziele nicht mehr erreicht oder gibt es starke Änderungen der Datengrundlage, wird dies dem*der Nutzer*in oder den Betroffenen kommuniziert und ggf. ein Prozess zur Aktualisierung oder zum kontrollierten Abschalten der KI-Anwendung angestoßen. Dabei ist sicherzustellen, dass im **Risikogebiet: Abfangen von Fehlern auf Modellebene (AF)** und im **Risikogebiet: Funktionale Sicherheit (FS)** alle relevanten Szenarien abgedeckt werden. Außerdem sind die Maßnahmen im **Risikogebiet: Beherrschung der Dynamik (BD)** in der Dimension Sicherheit zu berücksichtigen.

  Der Prozess sowie Art und Umfang der Überprüfungen werden dokumentiert.

7.5.4 Gesamtbewertung

[VE-R-BD-BW] Gesamtbewertung

Anforderung: Do

- Es wird dargelegt, dass ein Prozess zur regelmäßigen Überprüfung der KI-Anwendung aufgesetzt wurde, der die Kriterien in **[VE-R-BD-KR-01]** erfüllt.
- Sofern nicht alle in **[VE-R-BD-KR-01]** spezifizierten Anforderungen erfüllt werden, sind die Abweichungen zu dokumentieren. Dies gilt ebenfalls für nur teilerfüllte Anforderungen, bei denen etwa die Kriterien nicht, oder nicht immer, erreicht wurden.

Zusammenfassende Betrachtung

[VE-Z] Zusammenfassende Betrachtung der Dimension

Anforderung: Do

- Falls für diese Dimension ein mittlerer oder hoher Schutzbedarf besteht, ist eine Dokumentation über die verbleibenden Restrisiken zu erstellen. Zunächst werden die Restrisiken aus den verschiedenen Risikogebieten dieser Dimension zusammengefasst. Anschließend wird unter Berücksichtigung des Schutzbedarfs analysiert, ob die identifizierten Restrisiken insgesamt als vernachlässigbar, nicht vernachlässigbar (aber vertretbar) oder unvertretbar zu bewerten sind. Bei dieser Analyse sollten insbesondere die Auswirkungen von Maßnahmen aus der Dimension Sicherheit berücksichtigt werden, falls diese dazu beitragen, Fehler der KI-Komponente abzuschwächen oder zu verhindern. Das Ergebnis der Analyse ist zu erläutern. Außerdem ist, falls Trade-Offs

76 Bifet, A., Gavalda, R. (2007). Learning from time-changing data with adaptive windowing. In: Proceedings of the 2007 SIAM International Conference on Data Mining, SIAM, pp 443–448. https://doi.org/10.1137/1.9781611972771.42 (letzter Aufruf: 01.07.2021)

zwischen dem **Risikogebiet: Verlässlichkeit im Regelfall (RE)** und dem **Risikogebiet: Einschätzung von Unsicherheit (UN)** bestehen, die gewählte Priorisierung zu begründen.
- Falls potenziell negative Auswirkungen von Risiken oder Maßnahmen dieser Dimension auf andere Dimensionen, beispielsweise **Dimension: Fairness (FN)**, **Dimension: Transparenz (TR)** oder **Dimension: Sicherheit (SI)**, festgestellt wurden, sind diese zu dokumentieren.
- Es wird ein Fazit über die Dimension gezogen, welches insbesondere die Bewertung der Restrisiken enthält.

8. Dimension: Sicherheit (SI)

Beschreibung und Zielsetzung

Die Dimension Sicherheit adressiert Risiken bezüglich der Bereiche Funktionale Sicherheit und IT-Sicherheit, letztere in den Unterbereichen Integrität und Verfügbarkeit. Dabei bezeichnet die Funktionale Sicherheit (Bereich Safety) den Schutz vor Gefährdungen der Außenwelt vor funktionalem Versagen der KI-Anwendung. Eine typische Maßnahme ist beispielsweise die Installation von Airbags in einem Auto, um die Insassen bei einem Unfall zu schützen. IT-Sicherheit (Bereich Security) hingegen beschäftigt sich mit dem Schutz der KI-Anwendung vor seiner Umgebung, wie etwa äußeren Angriffen, die zu einer Veränderung oder Beeinträchtigung der Funktionalität führen können (Integrität). Hierzu eng verwandt sind Einbußen oder Verlust von Verfügbarkeit, bei der das System nicht oder nicht mehr im erforderlichen Maße antwortet oder reagiert. Verletzungen der Informationssicherheit, die beispielsweise zur Preisgabe von Informationen an unautorisierte Personen führen, werden gesondert in der **Dimension: Datenschutz (DS)** betrachtet.

Die High-Level Expert Group on AI (HLEG) hat abstrakte Sicherheitsziele für KI-Anwendungen definiert. Diese abstrakten Zielsetzungen (und darüberhinausgehende) sind jedoch weit von einer Operationalisierung etwa durch eine Norm entfernt. Umgekehrt existieren gerade im Bereich Sicherheit eine ganze Reihe von operativ überprüfbaren Spezifikationen und Normen, die jedoch keinen speziellen Bezug auf die Besonderheiten von KI-Anwendungen nehmen. Ziel der Dimension Sicherheit ist es, die Anforderungen aus bestehenden Normen, die unerlässlich für den Schutz vor Angriffen auf und Gefährdungen durch KI-Anwendungen sind, zusammenzuführen und mit weiteren KI-spezifischen Anforderungen zu ergänzen[77]. Dementsprechend beziehen sich die in diesem Dokument aufgeführten Risiken und Maßnahmen auf die KI-Anwendung bzw. stehen mit dieser im Zusammenhang. Sicherheitsrisiken der die KI-Komponente umgebenden klassischen Softwaremodule sind gemäß den vorhandenen Standards zu behandeln und werden in diesem Katalog nicht aufgeführt.

Gefährdungen im Bereich Sicherheit können sich als Funktionsausfall oder starke Funktionsänderung der KI-Komponente äußern. Solche Szenarien werden zwar auch in der **Dimension: Verlässlichkeit (VE)**, etwa in Form adversarialer Beispiele sowie Funktionsänderung durch *Concept Drift*, untersucht. Jedoch fokussieren die Kriterien und Maßnahmen in der **Dimension: Verlässlichkeit (VE)** auf die KI-Komponente, d. h. es werden dort nur solche Ursachen von Gefährdungen behandelt, die in der KI-Komponente liegen. Die Maßnahmen in der Dimension Sicherheit hingegen sind hauptsächlich auf die Einbettung bezogen, und nicht unmittelbar auf die KI-Komponente selbst anwendbar. Sie greifen insbesondere dann, wenn die Maßnahmen der **Dimension: Verlässlichkeit (VE)** keine vollumfängliche Abschwächung von Risiken gewährleisten können. So verweist beispielsweise das **Risikogebiet: Abfangen von Fehlern auf Modellebene (AF)** der Dimension Verlässlichkeit, das Detektionsmechanismen für stark abweichende Eingaben behandelt, für die an die Detektion anschließenden Mitigationsstrategien explizit auf das **Risikogebiet: Funktionale Sicherheit (FS)** der Dimension Sicherheit.

77 Die Darstellung in diesem Abschnitt ist stark angelehnt an das Kapitel »3.5 Sicherheit« des Whitepapers: Poretschkin, M., et al. (2019). Vertrauenswürdiger Einsatz von Künstlicher Intelligenz. Sankt Augustin: Fraunhofer-Institut für Intelligente Analyse und Informationssysteme IAIS. https://www.ki.nrw/wp-content/uploads/2020/03/Whitepaper_KI-Zertifizierung.pdf (letzter Aufruf: 18.06.2021)

Die Risikogebiete in der Dimension Sicherheit sind gegeben durch:

1. **Funktionale Sicherheit:** Dieses Risikogebiet behandelt das Risiko unbeabsichtigter Personen- oder Sachschäden, die durch Fehlfunktion bzw. Ausfall der KI-Anwendung infolge mangelhaften Designs der Einbettung begünstigt oder gar verursacht werden.

2. **Integrität und Verfügbarkeit:** Dieses Risikogebiet adressiert Risiken, die dadurch entstehen, dass die für den Betrieb der KI-Anwendung relevanten Daten verfälscht werden und dadurch die KI-Anwendung manipuliert wird und ggf. nicht mehr verfügbar ist.

3. **Beherrschung der Dynamik:** Dieses Risikogebiet behandelt Risiken, die dadurch entstehen, dass neue Gefährdungen der genannten Risikogebiete auftreten oder etablierte Verfahren an Effektivität verlieren.

Schutzbedarfsanalyse

In der Dimension Sicherheit werden zwei potenzielle Schadensszenarien betrachtet, vor deren Hintergrund der Schutzbedarf ermittelt wird: einerseits Personen- und Sachschäden sowie andererseits finanzielle Schäden.

Aus Sicht der Funktionalen Sicherheit haben insbesondere autonome Robotersysteme, wie beispielsweise selbstfahrende Fahrzeuge, die bei Unfällen einen Personen- oder Sachschaden verursachen können, einen erhöhten Schutzbedarf. Aus Sicht der IT-Sicherheit sind hingegen solche KI-Anwendungen besonders kritisch bei denen eine manipulierte oder inkorrekte Funktionsweise erhebliche finanzielle Schäden zur Folge haben könnte. Für Anwendungen mit Echtzeitbedarf, etwa zur automatisierten Abwicklung von Finanztransaktionen, kann auch die Gewährleistung niedriger Latenz von Bedeutung sein.

Die potenzielle Schadenshöhe ergibt sich daraus, ob eine Fehlfunktion der KI-Anwendung (z. B. eine inkorrekte Funktionsweise oder ein Ausfall) zu einem Personen- oder Sachschaden führen kann. Des Weiteren wird das Ausmaß von finanziellen Schäden berücksichtigt, die durch eine Fehlfunktion der KI-Anwendung verursacht werden können.

Der Schutzbedarf wird folgendermaßen kategorisiert:

Hoch	Es trifft mindestens einer der folgenden Punkte zu: ▪ Die KI-Anwendung interagiert mit Personen auf eine Art und Weise, dass diese bei einer Fehlfunktion verletzt werden können. ▪ Eine Fehlfunktion der KI-Anwendung (aufgrund von Fehlern, Ausfällen, Manipulation oder Attacken) kann zu einem sehr hohen finanziellen Schaden führen (z. B. durch Sachschäden). **Beispiel**: Eine KI-Anwendung, die zur Personen- und Objekterkennung in einem autonomen Fahrzeug eingesetzt wird. Im Fall einer inkorrekten Funktionsweise können sowohl Personen zu Schaden kommen als auch hohe finanzielle Kosten durch Sachbeschädigung entstehen. **Beispiel**: Eine KI-Diagnose-Anwendung, die Entscheidungen über die Art der medizinischen Behandlung von Personen trifft. Eine Manipulation der KI-Anwendung kann eine fehlerhafte Behandlung von Patient*innen zur Folge haben und sich dadurch gravierend auf deren Gesundheit auswirken.

Mittel	Durch die KI-Anwendung können Personen nicht direkt physisch verletzt werden. Eine Fehlfunktion der KI-Anwendung (aufgrund von Fehlern, Ausfällen, Manipulation oder Attacken) kann jedoch zu hohem finanziellen Schaden führen. **Beispiel:** Eine KI-gestützte Anwendung zum Warentransport in der Lagerhaltung kann, etwa bei unbeabsichtigtem Abladen von Waren in unzulässigen Bereichen oder Höhen, Waren beschädigen.
Gering	Eine Fehlfunktion der KI-Anwendung kann maximal zu mittlerem finanziellem Schaden führen. **Beispiel:** Eine KI-Anwendung, die zur Komposition von Musikstücken eingesetzt wird. Bei Ausfällen oder inkorrekter Funktionsweise ist kein finanzieller Schaden zu erwarten.

[SI-S] Dokumentation der Schutzbedarfsanalyse

Anforderung: Do

- Der Schutzbedarf der KI-Anwendung für die Dimension Sicherheit wird als *gering, mittel* oder *hoch* bestimmt. Die Festlegung der Kategorie *gering/mittel/hoch* wird unter Bezugnahme auf die oben angeführte Tabelle ausführlich begründet.

Falls der Schutzbedarf für die Dimension Sicherheit *gering* ist, so ist keine nähere Betrachtung der einzelnen Risikogebiete erforderlich. Dabei besteht die Ausnahme, dass, falls Maßnahmen im **Risikogebiet: Abfangen von Fehlern auf Modellebene (AF)** der Dimension Verlässlichkeit ergriffen werden, die im Zusammenhang mit dem **Risikogebiet: Funktionale Sicherheit (FS)** der Dimension Sicherheit stehen, das **Risikogebiet: Funktionale Sicherheit (FS)** in jedem Fall ordnungsgemäß behandelt wird. Wurde ein *mittlerer* oder *hoher* Schutzbedarf ermittelt, so muss jedes der folgenden Risikogebiete genauer untersucht werden.

8.1 Risikogebiet: Funktionale Sicherheit (FS)

Funktionale Sicherheit gehört zu dem Bereich von Sicherheit, der sich mit dem Schutz der Außenwelt vor durch die KI-Anwendung verursachten Gefährdungen befasst. Hierbei fokussiert der Prüfkatalog auf KI-spezifische Unfallrisiken[78] im Sinne von unbeabsichtigten und schädlichen Auswirkungen der KI-Anwendung, die durch mangelhaftes Design begünstigt oder gar verursacht werden. Das Risikogebiet Funktionale Sicherheit soll insbesondere das Risiko, dass etwa im Fall einer Fehlfunktion oder gar eines Ausfalls der KI-Anwendung Personen zu Schaden kommen oder materielle Schäden entstehen, auf ein akzeptables Maß reduzieren.

Da sich dieses Risikogebiet mit den Konsequenzen von Fehlfunktion bzw. Ausfall der KI-Anwendung auseinandersetzt, adressiert es eine Risikoklasse, die auch in der **Dimension: Verlässlichkeit (VE)** aufgegriffen wird. Die Maßnahmen in der Dimension Verlässlichkeit wirken unter anderem auf das Vermeiden von Fehlfunktionen (**Risikogebiet: Robustheit (RO)**) sowie das Erkennen von Fehlern (**Risikogebiet: Abfangen von Fehlern auf Modellebene (AF)**, **Risikogebiet: Einschätzung von Unsicherheit (UN)**) durch Methoden auf Ebene des ML-Modells hin. Funktionale Sicherheit schließt an diese Maßnahmen an, indem es das Erkennen von Fehlern bzw. Gefährdungen durch Methoden auf Einbettungsebene ergänzt und durch (z. T. klassische) Mitigationsstrategien auf Ebene der Einbettung gewährleistet, dass die Sicherheit im Fall erkannter Gefährdungen aufrechterhalten wird.

Typische Gefährdungen im Kontext der Funktionalen Sicherheit sind solche, die zu leichten bis schweren Verletzungen, Tod oder Sachschäden führen und durch Fehlverhalten oder Ausfall der KI-Anwendung entstehen. Eine mögliche Ursache dafür sind fehlerhafte oder schädliche Eingaben außerhalb des Anwendungsbereichs. In vielen Fällen können problematische Eingaben bereits auf Einbettungsebene erkannt werden. So kann beispielsweise eine Fehlermeldung implementiert werden, die anzeigt, wenn die Kamera einer bildverarbeitenden KI-Anwendung ausgefallen ist. Dadurch würde die Verarbeitung von Störsignalen durch die KI-Anwendung, die womöglich zu Fehlern führt, verhindert.

Für KI-Anwendungen, die komplexe Eingabedaten verarbeiten, reichen konventionelle Methoden jedoch oftmals nicht aus, um schädliche Eingaben zu erkennen. Dies ist insbesondere bei Open World Anwendungen wie etwa einer Spracherkennung der Fall, bei der konventionelle Methoden nicht in der Lage wären, zu erkennen, ob ein*e Nutzer*in die KI-Anwendung in der falschen Sprache bedient. Für solche Szenarien werden im **Risikogebiet: Abfangen von Fehlern auf Modellebene (AF)** der Dimension Verlässlichkeit Detektionsmechanismen auf Modellebene behandelt. So können problematische Eingabedaten, deren ungehinderte (Weiter-)Verarbeitung zu unvertretbarem Risiko in Hinblick auf einen Personen-, Sach- oder finanziellen Schaden führen würde, bereits durch die KI-Komponente abgefangen werden. Das Vorgehen im **Risikogebiet: Abfangen von Fehlern auf Modellebene (AF)** der Dimension Verlässlichkeit ist mit der Implementierung möglicher (zusätzlicher) konventioneller Funktionsüberwachungsmethoden in diesem Risikogebiet abzustimmen. Insbesondere sind, falls Fehler bzw. Gefährdungen auf Modellebene erkannt werden, Anschlussreaktionen (Mitigationsstrategien) einzurichten.

Ein weiterer Ansatzpunkt, um Fehler abzufangen bzw. Gefährdungen zu erkennen, ist die Durchführung eines *Sanity Checks* der Ausgaben der KI-Komponente, bevor diese durch die umgebenden Software-Module weiterverarbeitet werden. Im Fall einer KI-Anwendung zur Objekterkennung in Videoaufnahmen könnte man beispielsweise die aktuelle Segmentierungsmaske mit denen der vorigen Bilder abgleichen und auf starke Abweichungen überprüfen. Darüber hinaus sollte sichergestellt werden, dass nicht durch die Einbettung selbst, z. B. bei der Interpretation der Ausgabe der KI-Komponente, Fehler initiiert werden. Auch für den Fall, dass Fehler bzw. Gefährdungen auf Systemebene erkannt werden, müssen angemessene Anschlussreaktionen (Mitigationsstrategien) eingerichtet werden.

---

78 Das Konzept von Sicherheitsrisiken im Kontext von Künstlicher Intelligenz in diesem Risikogebiet orientiert sich an der Definition des Unfallrisikos im Paper: Amodei, D., et al. (2016). Concrete Problems in AI Safety. arXiv: 1606.06565 https://arxiv.org/abs/1606.06565 (letzter Aufruf: 30.06.2021)

Tritt ein Fehler auf bzw. wird eine Gefährdung erkannt, die lediglich einen punktuellen oder kurzzeitigen Ausfall unter beherrschbarem Risiko zur Folge hat, dann sollte die anschließende Mitigationsstrategie auf Fehlertoleranz hinwirken. Dazu setzt in der Regel eine Sicherheitsfunktion ein, die gewährleistet, dass die KI-Anwendung während des punktuellen Ausfalls abgesichert ist und gleichzeitig eine rudimentäre Funktionalität aufrechterhalten kann. Ist der Fehler behoben, so kann die KI-Anwendung ohne großen Aufwand wieder in den Normalbetrieb zurückkehren. Falls beispielsweise eine KI-Komponente für kontinuierliche Umfelderkennung ausfällt, so könnten die nicht gelieferten Daten für kurze Zeit aus den kürzlich erstellten Umfeldbeschreibungen approximiert werden, um den Ausfall kurzzeitig zu überbrücken.

Tritt hingegen ein Fehler auf bzw. wird eine Gefährdung erkannt, die zu einem sequenziellen Ausfall führen würde oder sogar den Verlust der Gesamtfunktionalität der KI-Anwendung unter unbeherrschbarem Risiko zur Folge hätte, so muss die KI-Anwendung in einen *Fail-Safe State* überführt werden. *Fail-Safe* bedeutet die Überführung in einen sicheren Zustand, wobei die oberste Priorität darin besteht, den Schaden auf ein möglichst geringes Ausmaß zu minimieren. Beim *Fail-Safe* wird unter Umständen die bestimmungsgemäße Funktion der KI-Anwendung deaktiviert und im Extremfall sogar die Zerstörung der KI-Anwendung oder des Gesamtsystems in Kauf genommen. Kann im obigen Beispiel der Umfelderkennung die KI-Anwendung etwa aufgrund von Sensorausfall nicht zeitnah wiederhergestellt werden, so könnte die KI-Anwendung durch Kontrollübergabe an den*die Nutzer*in (in diesem Fall wäre die KI-Anwendung dysfunktional) in einen *Fail-Safe State* überführt werden. Bei *Human-out-of-the-Loop*-Systemen, bei denen es keine*n direkte*n Nutzer*in gibt, ist der Übergang in den *Fail-Safe State* abhängig von Art und Betriebszustand der KI-Anwendung zu gestalten. Handelt es sich beispielsweise um eine KI-basierte Umfelderkennung in einem Staubsaugerroboter, so könnte ein *Fail-Safe State* bereits durch Abschalten der Anwendung erreicht werden. Ist die Umfelderkennung jedoch mit der Steuerung eines fahrenden unbemannten Transportfahrzeugs verbunden, so müsste ein *Fail-Safe State* unter Umständen die graduelle Verringerung der Geschwindigkeit und das Aussenden von Warnsignalen vorsehen.

Abgesehen von dem Szenario einer fehlerhaften Funktion oder gar eines Ausfalls der KI-Anwendung, können Sicherheitsrisiken auch bei korrektem Betrieb der KI-Anwendung entstehen, zum Beispiel aufgrund externer Faktoren. Im Beispiel einer KI-basierten Steuerung eines autonomen Fahrzeugs könnte plötzlich ein Mensch auf die Straße laufen. Wenn dieser Mensch korrekt durch die KI-Anwendung erkannt wird, kann trotzdem eine Gefährdung der Sicherheit bestehen, falls zum Beispiel der Bremsweg bis zum detektierten Menschen zu lang ist. Funktionale Sicherheit hat jedoch nur den Zweck, solche Risiken und Gefährdungen abzuschwächen, die von einer Fehlfunktion der KI-Anwendung ausgehen. So ist die Gefahr der Kollision mit durch die KI-Anwendung korrekt erkannten Menschen oder Objekten keine funktionale Gefahr und damit nicht im Fokus des Risikogebiets Funktionale Sicherheit.

8.1.1 Risikoanalyse und Zielvorgaben

[SI-R-FS-RI-01] Risikoanalyse und Zielvorgaben

Anforderung: Do

- **Risikoanalyse:** Gefährdungen bzw. potenzielle Schäden aufgrund von geringer Verlässlichkeit auf regulären Eingaben (siehe **[VE-R-RE-RI-01]**) oder einer Fehlfunktion bei aufkommenden Störungen (siehe **[VE-R-RO-RI-01]**) wurden bereits analysiert und deren Eintrittswahrscheinlichkeit abgeschätzt. Die genannten Risikoanalysen sind für die Zielsetzung dieses Risikogebiets relevant, müssen aber an dieser Stelle nicht wiederholt werden. Darüber hinaus spielt das Risiko von Fehlfunktionen der Einbettung, die in der **Dimension: Verlässlichkeit (VE)** nicht betrachtet wurden, für die Sicherheit der KI-Anwendung eine Rolle. Es ist zu analysieren, welche KI-spezifischen Faktoren zu einer Fehlfunktion der Einbettung führen können (dabei ist unter anderem zu klären, ob bestimmte Ausgaben der KI-Komponente bei der Weiterverarbeitung durch die Einbettung Fehler verursachen können), welche potenziellen Schäden daraus resultieren können und mit welcher Wahrscheinlichkeit diese auftreten. Bei der Untersuchung möglicher Ursachen für eine Fehlfunktion bzw. einen Ausfall der KI-Anwendung soll auch das (vorhersehbare) Nutzerverhalten berücksichtigt werden, darunter mindestens:
  - die vorhersehbare Fehlnutzung der Anwendung

  - der unerwartete Start der Anwendung
    (angelehnt an ISO-10218-1 4)
- **Gültigkeitsbereich für das Abfangen von Fehlern durch die Einbettung:** Es wird begründet dargelegt, welche Bereiche an Störungen oder Fehlern und welche Ein-/Ausgabeszenarien der KI-Komponente durch Methoden der Einbettung, d. h. auf Systemebene, abgefangen werden sollen.
  - Zum einen kann Fehlern bzw. Ausfällen der KI-Anwendung entgegengewirkt werden, indem Eingabedaten, die außerhalb der in **[VE-R-RO-RI-01]** festgelegten Anwendungsgrenze liegen, abgefangen werden.
    Beispiel: Die optische Umgebungserkennung einer autonom agierenden Einheit erkennt Glas nicht als Hindernis. Um Beschädigungen auf Einbettungsebene zu vermeiden, könnten Drucksensoren die Bewegung der Einheit bei Widerstand anhalten.
    Die Dokumentation, welche Eingabebereiche durch Methoden dieses Risikogebiets abgefangen werden sollen, ist im Einklang mit der Risikoanalyse in **[VE-R-AF-RI-01]** so zu gestalten, dass alle relevanten schädlichen Eingabebereiche abgedeckt sind. Insbesondere sollten auch solche Eingabedatenbereiche berücksichtigt werden, die nicht im Zusammenhang mit der KI-Anwendung oder ihrer Anwendungsgrenze stehen, die aber dennoch auftreten könnten.
  - Zum anderen können Fehler bzw. Störungen abgefangen werden, die sich erst durch die Einbettung ergeben. Jedoch adressiert dieser Prüfkatalog die KI-spezifischen Risiken einer KI-Anwendung und hat weder den Zweck noch den Anspruch, die klassische Funktionale Sicherheit oder auch Informationssicherheit vollständig abzubilden. Deshalb sind in dieser Dokumentation hinsichtlich der Einbettung nur solche Fehler bzw. Störungen der Einbettung zu betrachten, die in direktem Zusammenhang mit der KI-Komponente bzw. der Interpretation ihrer Ausgabe stehen. Darunter fällt beispielsweise nicht das Abfangen von Hardwarefehlern oder, dass das Überschreiten operativer Grenzen der Einbettung vermieden wird, sofern dies nicht im Zusammenhang mit der Funktion der KI-Komponente steht. Explizit zum Scope dieses Risikogebiets gehört hingegen das Abfangen von Ausgaben der KI-Komponente, die bei Weiterverarbeitung durch die Einbettung eine Gefährdung der Funktionalen Sicherheit zur Folge hätten. Ein Beispiel für solch eine Gefährdung ist, dass eine KI-Komponente zur Objektdetektion so viele Detektionen ausgibt, dass die Komponenten der Einbettung sie nicht in vertretbarer Zeit behandeln können. Basierend auf der Risikoanalyse ist festzulegen, welche Ausgabenszenarien durch die Einbettung abgefangen werden sollen.

  Es ist darzulegen, für welche Bereiche eine Fehlerdetektion als nicht erforderlich erachtet wird. Ferner ist zu begründen, dass durch die zur Detektion vorgesehenen Bereiche an Störungen, Fehlern, Ein- oder Ausgabeszenarien eine ausreichende Risikomitigation möglich ist.
- **Zielvorgaben:** Basierend auf der Risikoanalyse dieses Risikogebiets sowie unter Berücksichtigung von **[VE-R-RE-RI-01], [VE-R-RO-RI-01]** und **[VE-R-AF-RI-01]** werden qualitative Ziele für die Einbettung zur Absicherung der KI-Anwendung in Hinblick auf Funktionale Sicherheit formuliert. Dabei wird zum einen grob beschrieben, mit welcher Herangehensweise die im vorigen Abschnitt festgelegten Bereiche (an Fehlern, Störungen, Ein-/Ausgabeszenarien) abgefangen werden sollen. Zum anderen wird umrissen, in welchem Ausmaß und mit welcher grundlegenden Vorgehensweise die als relevant identifizierten Gefährdungen der Einbettung sowie die abgefangenen Fehler abgeschwächt bzw. behoben werden sollen. Hierbei ist auch auf potenzielle Anschlussreaktionen an die Detektionsmechanismen aus dem **Risikogebiet: Abfangen von Fehlern auf Modellebene (AF)** einzugehen. Insbesondere ist basierend auf den ermittelten Risiken grob zuzuordnen, in welchen Gefährdungsszenarien ein fehlertoleranter Ansatz (Aufrechterhaltung grundlegender Funktionalität) verfolgt wird, und ab wann der Übergang in einen *Fail-Safe State* (sicherer, aber dysfunktionaler Zustand) vorgesehen ist.

8.1.2 Kriterien zur Zielerreichung

Basierend auf den identifizierten Gefährdungen dieses Risikogebiets sollen entsprechende Absicherungsmaßnahmen getroffen werden, deren Ziel es ist, die hier adressierten Risiken auf ein vertretbares Maß zu senken. Um in der abschließenden Bewertung der Maßnahmen objektiv überprüfen zu können, ob diese Risiken erfolgreich mitigiert sind, müssen die in **[SI-R-FS-RI-01]** beschriebenen Zielvorgaben in quantitative Kriterien übersetzt werden. Dazu ist das angestrebte Risiko, das für den vorliegenden Anwendungskontext als vertretbar erachtet wird, zu spezifizieren. Außerdem sind Anforderungen an die Testdaten sowie die Mitigationsstrategien zu stellen.

[SI-R-FS-KR-01] Quantifizierung des vertretbaren Risikos

Anforderung: Do

- Es werden für den vorliegenden Anwendungskontext angemessene Kriterien aufgeführt, nach denen das Risiko potenzieller, durch Fehlfunktion der KI-Anwendung hervorgerufener Schäden bewertet wird. Diese Kriterien sollten mindestens umfassen:

  Schadenshöhe:
  – Art und Schwere möglicher Verletzungen
  – Anzahl der beeinträchtigten Personen
  – Kosten (Arbeitsausfall, Entschädigung und zusätzlicher Aufwand zur Behebung des Vorfalls sowie für Support)
  – Aufwand zur Behebung von Sachschäden

  Eintrittswahrscheinlichkeit:
  – unter Angabe möglicher Ursachen

- Für die aufgestellten Kriterien werden (evtl. pro Schadensszenario) Zielwerte festgelegt. Es ist nachvollziehbar zu argumentieren, weshalb das Risiko bei Einhaltung dieser Zielwerte ein vertretbares Ausmaß annimmt und mit ethisch-rechtlichen Rahmenbedingungen konform ist. Außerdem ist darzulegen, dass die hier festgelegten Kriterien die Zielvorgaben abbilden.

[SI-R-FS-KR-02] Anforderungen an die Testdaten

Anforderung: Do

- Angesichts der in diesem Risikogebiet betrachteten Aspekte, die über die **Dimension: Verlässlichkeit (VE)** hinausgehen, wie etwa
  – das Risiko einer Fehlfunktion der Einbettung,
  – das Abfangen von schädlichen Ein- oder Ausgaben der KI-Komponente durch die Einbettung,
  – Mitigationsstrategien der Einbettung,
  müssen die in **[VE-R-RO-KR-03]** und **[VE-R-AF-KR-01]** formulierten Kriterien zur Formalisierung und Quantifizierung der Abdeckung der Anwendungsgrenze sowie relevanter OOD-Bereiche für in diesem Risikogebiet zu verwendende Testdaten unter Umständen ergänzt werden. Sofern eine Quantifizierung begründet nicht möglich ist, wird auf eine qualitativ kategorische Auseinandersetzung zurückgegriffen, etwa auf Basis von *Zwicky Boxes*. Falls eine Ergänzung als nicht notwendig erachtet wird und die Testdatensätze aus der **Dimension: Verlässlichkeit (VE)** übernommen werden sollen, ist zu begründen, dass die o. g. Aspekte bereits ausreichend durch diese abgedeckt sind.

[SI-R-FS-KR-03] Vorhandensein von Mitigationsstrategien

Anforderung: Do

- Jedem gemäß **[SI-R-FS-RI-01]** durch Methoden der Einbettung abzufangendem (Fehler-)Bereich ist, vergleichbar zu **[VE-R-AF-KR-02]**, eine Mitigationsstrategie (aus **[SI-R-FS-MA-08]** oder **[SI-R-FS-MA-10]**) gegenübergestellt, die ergriffen werden soll, sofern das Detektionsverfahren anschlägt. Dies kann beispielsweise die Übergabe der Kontrolle an den*die Benutzer*in sein. Eine Strategie muss hierbei nicht für alle genannten Bereiche anwendbar sein, sondern kann sich auch auf spezifisch zu benennende Bereiche (oder auch nur Teilbereiche selbiger) beziehen. Es ist jedoch jeder (Teil-)Bereich mit mindestens einer Mitigationsstrategie zu belegen. Diese Zuordnung ist tabellarisch festgehalten.

[SI-R-FS-KR-04] Anforderungen an die Fehlererkennung

Anforderung: Do

- Für jeden der in **[SI-R-FS-RI-01]** genannten abzufangenden Bereiche werden, ausgehend von der Risikoanalyse sowie den gemäß **[SI-R-FS-KR-03]** zuzuordnenden Mitigationsstrategien, (falls möglich quantitative) Anforderungen an die Detektion festgehalten. Hierbei wird mindestens auf die folgenden Punkte eingegangen:
  - Verlässlichkeit (Festlegen einer Metrik, mit der die Performanz der Detektionsstrategie gemessen werden kann (siehe dazu **[VE-R-RE-KR-01]**), und eines Zielintervalls bzw. einer oberen Schranke, bis zu der ein Ausfall der Detektion vertretbar wäre)
  - Reaktionszeit (maximal zulässiger zeitlicher Versatz, abhängig von der anschließenden Mitigationsstrategie)
  - Einsatzschwelle (Sofern der Übergang zwischen »robustem« Verhalten und Ausfall der KI-Komponente im Eingaberaum kontinuierlich ist, wird begründet ein Schwellwert festgelegt, ab dem die Detektion anschlagen soll.)
  - Generalisierungsfähigkeit (Sofern die Abdeckung in **[SI-R-FS-KR-02]** qualitativ statt quantitativ ist, wird die Anforderung der Generalisierung bzw. Extrapolation für die Methode diskutiert.)
  - (Falls zutreffend,) spezifische weitere Anforderungen, die sich aus der jeweiligen, gemäß **[SI-R-FS-KR-03]** festzulegenden, anschließenden Mitigationsstrategie ergeben

[SI-R-FS-KR-05] Anforderungen an die Mitigationsstrategien mit dem Ziel Fehlertoleranz

Anforderung: Do

- Eine Mitigationsstrategie mit dem Ziel Fehlertoleranz muss für die sicherheitsrelevanten Teile der KI-Anwendung folgende Eigenschaften erreichen (charakteristische Merkmale):
  - Ein einzelner Fehler oder ein punktueller Ausfall der KI-Komponente führt nicht zum Verlust der Sicherheit.
  - Wenn verhältnismäßig und vernünftigerweise durchführbar, wird jeder einzelne Fehler dem*der Nutzer*in und bestenfalls dem*der Entwickler*in oder dem*der Betreiber*in gemeldet. Hierbei ist sicherzustellen, dass Fehlermeldungen, die sensible Daten enthalten, entsprechend geschützt sind (z. B. personenbezogene Informationen, vgl. **[SI-R-IV-MA-07]** und **[SI-R-IV-MA-11]**).
  - Bei Auftreten des einzelnen Fehlers wird eine grundlegende Funktionalität im sicheren Zustand aufrechterhalten, bis der erkannte Fehler behoben ist. Kann der erkannte Fehler nicht behoben und der sichere Zustand nicht länger aufrechterhalten werden, so muss ein Übergang in einen *Fail-Safe State* eingeleitet werden (siehe **[SI-R-FS-MA-10]**).
    (angelehnt an: ISO-10218-1 5.4)
- Pro vorgesehener Mitigationsstrategie mit dem Ziel Fehlertoleranz sind die drei genannten charakteristischen Merkmale ausgehend von den in **[SI-R-FS-RI-01]** ermittelten Risiken zu (falls möglich quantitativen) anwendungsspezifischen Kriterien auszuformulieren. Hierbei wird mindestens auf die nachfolgenden Punkte eingegangen, und es werden je nach Anwendungskontext weitere Anforderungen ergänzt:
  - Fehlerszenarien bzw. (falls möglich) Schwellwerte oder qualitative Kriterien, bei denen die Mitigationsstrategie greifen soll
  - Art und Umfang der Funktionalitäten, die durch Einsetzen der Mitigationsstrategie aufrechterhalten werden sollen
  - Verlässlichkeit (Festlegen einer Metrik und eines Zielwertes, mit der die Performanz der Mitigationsstrategie beurteilt werden kann, siehe hierzu **[VE-R-RE-KR-01]**)
  - U. U. Länge des Zeitfensters, in dem die Verlässlichkeit mindestens auf diesem Level gehalten werden muss
  - Reaktionszeit (maximal zulässiger zeitlicher Versatz zum Einsetzen der Mitigationsstrategie)
  - Fehlerszenarien bzw. (falls möglich) Schwellwerte oder qualitative Kriterien, ab denen die Mitigationsstrategie in einen *Fail-Safe State* übergehen muss (dabei sind u. a. die operativen Grenzen der Einbettung zu berücksichtigen).

[SI-R-FS-KR-06] Anforderungen an die Mitigationsstrategien mit dem Ziel *Fail-Safe*

Anforderung: Do

- Ausgehend von dem vorliegenden Anwendungskontext und den damit verbundenen, in **[SI-R-FS-RI-01]** ermittelten Risiken, werden für jede gemäß **[SI-R-FS-KR-03]** zuzuordnende Mitigationsstrategie mit dem Ziel *Fail-Safe* (falls möglich quantitative) Anforderungen festgehalten. Hierbei wird mindestens auf die folgenden Punkte eingegangen:
  - Fehlerszenarien bzw. (falls möglich) zugehörige Schwellwerte oder qualitative Kriterien, ab denen die Mitigationsstrategie mit dem Ziel *Fail-Safe* greifen soll. Dabei sind operative Grenzen der Einbettung, z. B. maximale Drehzahl bei einer Motorsteuerung, zu berücksichtigen. Außerdem zählen hierzu auch die Szenarien, in denen eine *Fail-Safe*-Strategie im Anschluss an eine Mitigationsstrategie mit dem Ziel Fehlertoleranz ausgelöst wird (vgl. **[SI-R-FS-KR-05]**).
  - Reaktionszeit (maximal zulässiger zeitlicher Versatz bis zum Einsetzen der Mitigationsstrategie)
  - Beschreibung des *Fail-Safe States*, der im Optimalfall erreicht werden soll
  - (Falls zutreffend) Verlässlichkeit der *Fail-Safe*-Strategie, d. h. die Zuverlässigkeit, mit der der *Fail-Safe State* erreicht wird
  - Vertretbarer maximaler Schaden, der bei Übergang in den *Fail-Safe State* in Kauf genommen werden darf

8.1.3 Maßnahmen

Die Maßnahmen in diesem Risikogebiet umfassen Methoden und Prüfungen der Einbettung, um Fehler bzw. einen Ausfall der KI-Anwendung abzufangen, (im Sinne der Fehlertoleranz) zu überbrücken sowie ggf. einen *Fail-Safe*-Zustand herzustellen.

[SI-R-FS-MA-01] Sicherheitsrichtlinien und Nutzungsanweisungen

Anforderungen: Do | Pr

- Es werden Sicherheitsziele für die Funktionale Sicherheit aus den Unternehmenszielen, Geschäftsprozessen, relevanten Gesetzen, Verordnungen und möglichen Gefährdungen abgeleitet und dokumentiert. Diese Sicherheitsrichtlinie enthält außerdem strategische Vorgaben, wie diese Ziele erreicht werden sollen.
- Basierend auf der Sicherheitsrichtlinie werden in einheitlicher Struktur Anweisungen für
  - die sichere Nutzung der Anwendung und
  - die Entwicklung der Anwendung

  festgehalten. Es sind Maßnahmen beschrieben, die dafür sorgen, dass alle Nutzer*innen diese zur Kenntnis nehmen.
- Es ist ein Prozess etabliert und dokumentiert, der Nutzer*innen und Entwickler*innen der Anwendung für Risiken bezüglich Funktionaler Sicherheit sensibilisiert und auf den korrekten Umgang in Bezug auf die Sicherheit von Daten, Modell und Einbettung hinweist.
  (angelehnt an BSI-C5 SA-01)

8.1.3.1 Daten

[SI-R-FS-MA-02] Szenarienabdeckung

Anforderung: Do

- Es liegt eine Dokumentation vor, in der beschrieben wird, welche Testdaten zur Überprüfung von Maßnahmen in diesem Risikogebiet verwendet werden. Dabei wird nachvollziehbar erläutert, dass die Testdaten genügend potenzielle Unfallszenarien und kritische Situationen beinhalten und somit die in **[SI-R-FS-KR-02]** definierten Kriterien erfüllen. Hierbei kann ggf. auf Dokumentationen aus der **Dimension: Verlässlichkeit (VE)** verwiesen werden.

8.1.3.2 KI-Komponente

[SI-R-FS-MA-03] Beitrag der KI-Komponente

Anforderung: Do

- Es wird dargelegt, inwiefern Architektur und Design der KI-Komponente zur Funktionalen Sicherheit der KI-Anwendung, insbesondere zur Vermeidung von Unfällen, beitragen. Falls zutreffend, ist zu beschreiben, ob die Lernfunktion (bzw. die Erstellung der Lernfunktion) Unfälle und Verletzungen mit negativem Feedback berücksichtigt. Falls dies bereits in der **Dimension: Verlässlichkeit (VE)** beschrieben wurde, kann dafür auf die entsprechenden Dokumentationen verwiesen werden.

8.1.3.3 Einbettung

[SI-R-FS-MA-04] Design der Einbettung

Anforderung: Do

- Es liegt eine Dokumentation vor, in der festgehalten ist, inwiefern Design und Architektur der Einbettung (z. B. durch redundante Auslegung oder durch die Einbindung klassischer Assistenzsysteme) zur Vermeidung oder auch Überbrückung einer Fehlfunktion und damit zur Stärkung der Funktionalen Sicherheit beitragen.
- Da dieser Prüfkatalog in erster Linie die KI-spezifischen Risiken der KI-Anwendung adressiert und weder den Zweck noch den Anspruch hat, bestehende klassische Standards (zur Funktionalen Sicherheit, Produktsicherheit, Informationssicherheit, etc.) vollständig abzubilden, sollte die Dokumentation jene Aspekte der Einbettung fokussieren, die in Zusammenhang mit der Weiterverarbeitung bzw. Interpretation der Ausgabe der KI-Komponente stehen. Zur Behandlung der KI-unspezifischen Risiken sind die gängigen klassischen Normen und Standards hinzuzuziehen.

[SI-R-FS-MA-05] Abfangen schädlicher Eingabedaten

Anforderungen: Do | Te

- Es wird nachvollziehbar erläutert, durch welche Methoden auf Ebene der Einbettung schädliche Eingaben außerhalb der Anwendungsgrenze, deren Verarbeitung ein unvertretbares Sicherheitsrisiko darstellen würde, erkannt werden. Ein typischer Fehlermodus, der durch klassische Methoden erkannt werden kann, ist etwa die Fehlfunktion eines Sensors. Im Fall von Bilddaten könnten schädliche Eingaben beispielsweise durch Messen toter Pixel erkannt und abgefangen werden.
  - Falls zum Abfangen von schädlichen Eingaben Methoden auf Einbettungsebene ergänzend zu Detektionsmechanismen auf Modellebene implementiert werden, so sind diese mit dem Vorgehen im **Risikogebiet: Abfangen von Fehlern auf Modellebene (AF)** der Dimension Verlässlichkeit abzustimmen. Ferner ist darzulegen, dass alle Eingabebereiche, für die in **[VE-R-AF-RI-01]** eine Detektion aus der Einbettung heraus vorgesehen wird, tatsächlich durch die hier vorgestellten Maßnahmen wirksam abgedeckt sind.
- Die Wirksamkeit dieser Maßnahme zum Abfangen von Fehlern bzw. Erkennen von Gefährdungen wird in geeigneten Tests nachgewiesen. Die Tests sind zu dokumentieren und deren Wahl zu begründen. Falls kein separater Test dieser Maßnahme vorliegt, ist darzulegen, dass diese bereits durch **[SI-R-FS-MA-09]**, **[SI-R-FS-MA-11]** oder den abschließenden Realtest **[SI-R-FS-MA-13]** ausreichend untersucht wurde.
- Es ist zu erläutern, inwiefern die beschriebene(n) Methode(n) zur Erfüllung des Kriteriums **[SI-R-FS-KR-04]** beitragen.

[SI-R-FS-MA-06] Abfangen von Fehlern bei Interpretation der Ausgabe der KI-Komponente

Anforderungen: Do | Te

- Es wird nachvollziehbar erläutert, durch welche Methoden auf Ebene der Einbettung Ausgaben der KI-Komponente, deren Weiterverarbeitung/Interpretation durch die Einbettung ein unvertretbares Sicherheitsrisiko darstellen würde, erkannt werden. Beispielsweise könnte ein KI-basiertes System zur Kollisionsvermeidung die Interpretation der KI-Komponente zusätzlich dadurch absichern, dass unabhängig von der Ausgabe

der KI-Komponente Ergebnisse von Abstands- und Geschwindigkeitssensoren abgeglichen werden oder die Ausgabe der KI-Komponente auf (bspw. zeitliche) Konsistenz geprüft wird.
- Die Wirksamkeit dieser Maßnahmen auf Einbettungsebene zum Abfangen von Fehlern bzw. Erkennen von Gefährdungen wird in geeigneten Tests nachgewiesen. Die Tests sind zu dokumentieren und deren Wahl zu begründen. Falls kein separater Test der Maßnahme stattfindet, ist darzulegen, dass diese bereits durch **[SI-R-FS-MA-09]**, **[SI-R-FS-MA-11]** oder den abschließenden Realtest **[SI-R-FS-MA-13]** ausreichend untersucht wurde.
- Es ist zu erläutern, inwiefern die beschriebene(n) Methode(n) zur Erfüllung des Kriteriums **[SI-R-FS-KR-04]** beitragen.

[SI-R-FS-MA-07] Wahl der Mitigationsstrategie

Nachdem ein Fehler abgefangen bzw. ein Gefährdungsszenario durch Methoden der Einbettung oder das Modell erkannt wurde, muss eine Mitigationsstrategie einsetzen, die die Sicherheit in dieser Ausnahmesituation gewährleistet. Ob die Mitigationsstrategie zur Fehlertoleranz der KI-Anwendung im Sinne einer kurzzeitigen Überbrückung der Ausnahmesituation bei Aufrechterhaltung grundlegender Funktionalität beiträgt, oder ob die Mitigationsstrategie in einen *Fail-Safe State* überführt, hängt von der Art des Fehlers sowie dem Betriebs- und Umgebungszustand der KI-Anwendung und der damit verbundenen Risikolage ab.
Beispiel: Eine Objektdetektion im automatisiert fahrenden Auto kann Detektionen, deren erwartete Unsicherheit derart hoch ist, dass die Position nicht mehr zuverlässig bestimmt werden kann, kurzfristig übergehen, wenn der zuletzt (sicher) bekannte Abstand zum Objekt angesichts der Geschwindigkeit des Autos hinreichend groß ist. Andernfalls sollte es z. B. direkt abbremsen.

Anforderung: Do

- Es ist zu dokumentieren, nach welcher Vorgehensweise (z. B. »by design« oder situationsabhängig) und nach welchen Kriterien festgelegt wird, durch welche anschließende Mitigationsstrategie ein (durch die Einbettung oder die KI-Komponente selbst) abgefangener Fehler bzw. eine erkannte Gefährdung behandelt wird.
  - Dabei ist unter Einbezug der Risikoanalysen **[SI-R-FS-RI-01]** und **[VE-R-AF-RI-01]** darzulegen, dass die Risiken und Konsequenzen eines fehlertoleranten Umgangs ausreichend berücksichtigt wurden.
  - Außerdem ist darzulegen, dass die Vorgehensweise zur Wahl der anschließenden Mitigationsstrategie konsistent mit den Zielvorgaben in **[SI-R-FS-RI-01]** ist.
- Es ist zu begründen, dass die beschriebene Vorgehensweise zur Wahl der Mitigationsstrategie dem vorliegenden Anwendungskontext angemessen ist und zur Funktionalen Sicherheit der KI-Anwendung beiträgt.
- Gemäß **[SI-R-FS-KR-03]** liegt eine tabellarische Übersicht vor, die pro vorhandenem Detektionsmechanismus (auf Modellebene und auf Ebene der Einbettung) die anschließende Mitigationsstrategie einschließlich ihres Ziels (d. h. Fehlertoleranz oder *Fail-Safe*) aufzeigt.

[SI-R-FS-MA-08] Mitigationsstrategien mit dem Ziel Fehlertoleranz

Fehlertoleranz bezeichnet die Eigenschaft der KI-Anwendung, ihre Funktionalität (zumindest grundlegend) auch dann aufrecht zu erhalten, falls eine Fehlfunktion oder ein Ausfall der KI-Komponente eintritt. Um Fehlertoleranz zu erreichen, muss die KI-Anwendung über Mitigationsstrategien verfügen, die beispielsweise mit schädlichen Eingaben umgehen oder kurzzeitig einen Ausfall überbrücken, bis der Fehler behoben ist.

Anforderungen: Do | Te

- Es liegt eine Dokumentation vor, die die Fehlertoleranz der KI-Anwendung eindeutig beschreibt. Dazu wird ausgehend von der tabellarischen Übersicht aus **[SI-R-FS-MA-07]** im Detail erläutert, über welche Mitigationsstrategien mit dem Ziel Fehlertoleranz die KI-Anwendung verfügt.
  - Falls zutreffend (und nicht schon als Detektionsmechanismus auf Modellebene behandelt), sind außerdem Anschlussreaktionen bezüglich einer Unsicherheitsschätzung (siehe **Risikogebiet: Einschätzung von Unsicherheit (UN)** der Dimension Verlässlichkeit) festzulegen und zu dokumentieren.

- Die Wirksamkeit der Mitigationsstrategien mit dem Ziel Fehlertoleranz in den für sie gemäß **[SI-R-FS-KR-05]** vorgesehenen Einsatzszenarien wird durch entsprechende Tests nachgewiesen. Die Tests sind zu dokumentieren und deren Wahl zu begründen. Falls kein separater Test der Mitigationsstrategien vorliegt, ist darzulegen, dass diese bereits durch **[SI-R-FS-MA-09]** oder den umfassenden Realtest **[SI-R-FS-MA-13]** ausreichend untersucht wurden.
- Es wird dargelegt, dass die vorhandenen Mitigationsstrategien mit dem Ziel Fehlertoleranz die in **[SI-R-FS-KR-05]** festgelegten Kriterien erfüllen.
- Außerdem wird ausführlich begründet, dass die erzielte Fehlertoleranz für den vorliegenden Anwendungskontext ausreichend und angemessen ist.

[SI-R-FS-MA-09] Test der Fehlertoleranz inklusive ihrer Detektionsmechanismen

Anforderungen: Do | Te

- Es werden Tests der KI-Anwendung in unerwarteten Situationen und in verschiedenen Gefährdungsumgebungen, in denen die KI-Anwendung fehlertolerant sein soll, durchgeführt und dokumentiert.
  - Dabei werden Detektionsmechanismen auf Modell- und Einbettungsebene für schädliche Eingaben gezielt ausgelöst, um daran anschließende Mitigationsstrategien mit dem Ziel Fehlertoleranz (vgl. **[SI-R-FS-MA-08]**) zu testen. Dazu enthalten die Testdaten mindestens diejenigen Szenarien des Datensatzes aus **[SI-R-FS-MA-02]** und **[VE-R-AF-MA-01]** (und falls zutreffend auch **[VE-R-AF-MA-02]**), für die die KI-Anwendung fehlertolerant sein soll.
  - Die Testdaten sind so zu erweitern, dass auch solche Ausgaben der KI-Komponente provoziert werden, die auf Einbettungsebene zum Zwecke der Fehlertoleranz abgefangen werden sollen.
- Die Wahl der Testdatensätze ist zu begründen. Zudem ist darzulegen, dass diese den Kriterien **[SI-R-FS-KR-02]** genügen.
- Es liegt eine Dokumentation vor, in der begründet ist, dass die durchgeführten Tests und deren Ergebnisse ausreichen, um der KI-Anwendung ein angemessenes fehlertolerantes Verhalten in unerwarteten Situationen und in Gefährdungsumgebungen zu attestieren.

[SI-R-FS-MA-10] Mitigationsstrategien mit dem Ziel *Fail-Safe*

Anforderungen: Do | Te

- Es liegt eine Dokumentation vor, die die Mitigationsstrategien der KI-Anwendung mit dem Ziel *Fail-Safe* ausgehend von der tabellarischen Übersicht aus **[SI-R-FS-MA-07]** im Detail erläutert. Zudem wird für jede Mitigationsstrategie der dadurch herzustellende *Fail-Safe State* beschrieben. Dabei werden unter anderem die folgenden Aspekte adressiert:
  - Falls eine *Fail-Safe*-Strategie in einem Einsatzszenario absehbar Schäden verursacht, so ist darzulegen, dass eine Abwägung verschiedener Strategien stattgefunden hat und die gewählte Mitigationsstrategie den Schaden minimiert.
  - Bei Mitigationsstrategien mit dem Ziel *Fail-Safe*, die den*die Nutzer*in einbinden (z. B. durch Kontrollübergabe an den*die Nutzer*in), ist darzulegen, dass der*die Nutzer*in über diese Möglichkeit informiert ist und unterrichtet wurde, wie die Person in solch einer Situation zu agieren hat. Wurde dies bereits in der **Dimension: Autonomie und Kontrolle (AK)** oder in **[SI-R-FS-MA-12]** behandelt, oder durch **[SI-R-FS-MA-01]** erfüllt, so ist auf die entsprechende Stelle zu verweisen.
- Die Wirksamkeit der Mitigationsstrategien mit dem Ziel *Fail-Safe* in den für sie gemäß **[SI-R-FS-KR-06]** vorgesehenen Einsatzszenarien wird durch entsprechende Tests nachgewiesen. Die Tests sind zu dokumentieren und deren Wahl zu begründen. Falls kein separater Test der Mitigationsstrategien vorliegt, ist darzulegen, dass diese bereits durch **[SI-R-FS-MA-11]** oder den umfassenden Realtest **[SI-R-FS-MA-13]** ausreichend untersucht wurden.
- Für jede vorhandene *Fail-Safe*-Strategie wird ausführlich begründet, dass sie für die Anwendungsszenarien, in denen sie gemäß **[SI-R-FS-KR-06]** (und **[SI-R-FS-KR-05])** ausgelöst werden soll, verhältnismäßig und angemessen ist.

[SI-R-FS-MA-11] Test der Fail-Safe-Strategien inklusive ihrer Detektionsmechanismen

Anforderungen: Do | Te

- Die KI-Anwendung wird in Tests verschiedenen Gefährdungen ausgesetzt, bei deren Eintreten ein *Fail-Safe* eingeleitet werden muss. Die Durchführung der Tests ist zu dokumentieren.
  - In den Tests werden solche Detektionsmechanismen auf Modell- und Einbettungsebene gezielt ausgelöst, deren anschließende Mitigationsstrategie in einen *Fail-Safe State* überführen soll.
    - Dazu enthalten die Testdaten mindestens diejenigen Szenarien des Datensatzes aus **[SI-R-FS-MA-02]**, **[VE-R-AF-MA-01]** (und falls zutreffend auch **[VE-R-AF-MA-02]**), an die ein *Fail-Safe* anschließen soll.
    - Die Testdaten sind so zu erweitern, dass auch die Detektionsmechanismen auf Ebene der Einbettung, die schädliche Ausgaben der KI-Komponente abfangen und an die ein *Fail-Safe* anschließen muss, abgedeckt werden.
  - Darüber hinaus sind, sofern existent, gezielt jene Szenarien zu provozieren, bei denen die KI-Anwendung von einem fehlertoleranten Verhalten in ein *Fail-Safe* übergehen muss.
- Die Wahl der Testdatensätze ist zu begründen. Zudem ist darzulegen, dass diese den Kriterien **[SI-R-FS-KR-02]** genügen.
- Es liegt eine Dokumentation vor, in der begründet ist, dass die durchgeführten Tests und deren Ergebnisse ausreichen, um der KI-Anwendung zu attestieren, dass sie in allen erforderlichen Situationen angemessen in einen *Fail-Safe State* überleitet.

[SI-R-FS-MA-12] Möglichkeit des menschlichen Eingriffs

Anforderung: Do

- Es liegt eine Dokumentation darüber vor, inwieweit der menschliche Eingriff in die Funktionsweise oder den Betrieb der KI-Anwendung notwendig und möglich ist, um einen Unfall zu vermeiden. Falls dort bereits thematisiert, kann auf die Mitigationsstrategien (siehe **[SI-R-FS-MA-08]** und **[SI-R-FS-MA-10]**) oder auf die **Dimension: Autonomie und Kontrolle (AK)** (siehe **[AK-R-GE-MA-02]**) verwiesen werden.
- Wurde für die KI-Anwendung ein hoher Schutzbedarf in der Dimension Sicherheit festgestellt, so ist darzulegen, dass sie über eine ausgewiesene Stoppfunktion verfügt, die
  - Vorrang vor allen anderen Funktionen der Anwendung hat,
  - das Stillsetzen aller von gesteuerten Teilen ausgehenden Gefährdungen bewirkt,
  - die Möglichkeit bietet, Gefährdungen zu beherrschen, die von der Anwendung ausgehen,
  - bis zum Rücksetzen aktiv bleibt und
  - nur durch explizite Bestätigung zurückgesetzt werden kann.
    (angelehnt an: ISO-10218-1 5.5)

  Sofern bei hohem Schutzbedarf die Bedienung der Stoppfunktion auf eine bestimmte Personengruppe eingeschränkt ist (bzw. gewisse Personengruppen von deren Bedienung ausgeschlossen sind), ist dies in Hinblick auf die Risiken ausführlich zu begründen.
- Da spontane Abschaltungen sowie andere Formen des menschlichen Eingriffs unter Umständen Risiken bergen können (z. B. kann eine Gefährdung aufgrund von Trägheit weiterhin bestehen), ist außerdem zu dokumentieren, dass potenzielle Nutzer*innen über den angemessenen Umgang mit den Eingriffsmöglichkeiten und daraus potenziell resultierende Konsequenzen unterrichtet wurden. Falls dort bereits thematisiert, kann auf **[SI-R-FS-MA-01]** oder auf die **Dimension: Autonomie und Kontrolle (AK)** verwiesen werden.

[SI-R-FS-MA-13] Umfassender Realtest

Anforderung: Do | Te

- Vor Inbetriebnahme der KI-Anwendung wird diese bereits im Gesamtsystem installiert und funktionstüchtig, einem Realtest unterzogen. Dieser deckt alle in diesem Risikogebiet behandelten Sicherheitsfunktionen und deren Verkettung ab. Die KI-Anwendung wird mit allen als relevant identifizierten Triggern von Fehlfunktionen konfrontiert, die zunächst durch die Anwendung detektiert werden müssen und anschließend eine Mitigationsstrategie mit dem Ziel Fehlertoleranz oder den Übergang in einen *Fail-Safe State* auslösen sollen. Die Durchführung der Tests wird dokumentiert.

- Die Wahl der Testszenarien ist zu dokumentieren und zu begründen. Zudem ist darzulegen, dass diese den Kriterien **[SI-R-FS-KR-02]** genügen. Ggf. kann auf entsprechende Stellen in der **Dimension: Verlässlichkeit (VE)** verwiesen werden. Es wird dargelegt, dass der Test und die erzielten Testergebnisse ausreichen, um der KI-Anwendung ein gemäß **[SI-R-FS-KR-01]** angemessenes Verhalten in Gefährdungsumgebungen und Gefahrensituationen zu attestieren.
- Tests von Maßnahmen der Funktionalen Sicherheit, die KI-unspezifisch sind und dementsprechend nicht in diesem Risikogebiet behandelt werden, sind gesondert und entsprechend der bestehenden Normen und Standards durchzuführen.

8.1.3.4 Maßnahmen für den Betrieb

[SI-R-FS-MA-14] Unfall-Behandlung

Anforderungen: Do | Pr

- Die Art und der Verlauf von im Zusammenhang mit der KI-Anwendung eintretenden Unfällen wird protokolliert und es wird festgehalten, wie die KI-Anwendung mit der jeweiligen Unfallsituation umgeht. Die eingetretenen Unfälle und deren Ursache werden analysiert.
- Es wird kontinuierlich überprüft, ob die Funktionsweise der KI-Anwendung in einer Unfallsituation den in **[SI-R-FS-RI-01]** festgelegten Zielvorgaben entspricht, und ob die Maßnahmen in diesem Risikogebiet zur Einhaltung der in **[SI-R-FS-KR-01]** festgelegten Kriterien in ausreichendem Maße beitragen. Werden Abweichungen festgestellt, so müssen Anpassungen an den Sicherheitsmaßnahmen oder an der KI-Komponente selbst (siehe **Dimension: Verlässlichkeit (VE)**) vorgenommen werden. Diese sind zu dokumentieren.

8.1.4 Gesamtbewertung

[SI-R-FS-BW] Gesamtbewertung

Anforderung: Do

- Unter Bezugnahme auf die durchgeführten und dokumentierten Tests wird nachvollziehbar dargelegt, dass gemäß **[SI-R-FS-KR-01]** ein Betrieb der KI-Anwendung unter vertretbarem Unfallrisiko gewährleistet ist, insbesondere auch hinsichtlich unbekannter Eingangsdaten, sofern dies erforderlich ist.
- Sollten die Maßnahmen in diesem Risikogebiet nicht realisierbar sein oder nicht ausreichen, um die Kriterien **[SI-R-FS-KR-01]** bis **[SI-R-FS-KR-06]** zu erfüllen, so ist dies zu dokumentieren. Die hier nicht behandelbaren Problemstellungen können in der dimensionsübergreifenden Gesamtbewertung abgewogen werden.
- Im Fall, dass Sicherheitsmaßnahmen aus diesem Risikogebiet und Detektionsmaßnahmen aus dem **Risikogebiet: Abfangen von Fehlern auf Modellebene (AF)** der Dimension Verlässlichkeit einander ergänzen, ist dokumentiert, dass die Tests nur eine schwache Korrelation untereinander aufweisen, sodass das Risiko eines gleichzeitigen oder untereinander bedingten Versagens als beherrschbar angesehen werden kann.
- Darüber hinaus sind Standards bzw. Normen der Funktionalen Sicherheit zu dokumentieren, die zusätzlich zu diesem Prüfkatalog zur Prüfung der KI-Anwendung herangezogen werden.

8.2 Risikogebiet: Integrität und Verfügbarkeit (IV)

Integrität und Verfügbarkeit sind Schutzziele der klassischen IT-Sicherheit. Sie werden in diesem KI-Prüfkatalog erneut aufgegriffen, da einerseits bestehende Risiken in Bezug auf diese Schutzziele durch den Einsatz Maschineller Lernverfahren gesteigert werden; andererseits ergeben sich neuartige Risiken, etwa weil Daten für KI-Technologien einen sensibleren Angriffsvektor darstellen, als es bei klassischen IT-Systemen der Fall ist. Insbesondere ergibt sich aus der Tatsache, dass KI-Anwendungen datengetriebene IT-Systeme sind, eine stärkere Überschneidung zwischen den KI-spezifischen Risiken hinsichtlich Integrität und Verfügbarkeit, weshalb diese Schutzziele zu einem Risikogebiet zusammengefasst werden.

Im Kontext der Informationssicherheit ist Integrität gleichbedeutend mit Unversehrtheit in dem Sinne, dass keine unautorisierten bzw. unbeabsichtigten Änderungen vorgenommen werden. In diesem Risikogebiet wird die Integrität der KI-Anwendung betrachtet. Aufgrund der Natur Maschineller Lernverfahren haben Daten einen essenziellen Einfluss auf die Qualität und Funktionalität der KI-Anwendung und bilden somit eine Angriffsfläche für deren Integrität. Verletzungen der Integrität können verschiedene Ausmaße annehmen.

Zum einen kann die Integrität einer KI-Anwendung punktuell untergraben werden. Insbesondere verletzen adversariale Beispiele die Integrität der KI-Anwendung, wenn sie gezielt von Angreifer*innen erstellt wurden, um etwa eine bestimmte Ausgabe zu erzwingen. Solche Angriffe werden als adversariale Attacken bezeichnet. Zwar nutzen adversariale Attacken Schwachstellen des ML-Modells aus und sind in ihrer Ursache der **Dimension: Verlässlichkeit (VE)** zuzuordnen. Jedoch stellt die Angreifbarkeit Neuronaler Netze ein Sicherheitsrisiko dar, das zusätzlich zu den Maßnahmen im **Risikogebiet: Robustheit (RO)**, die sich vorrangig auf das Design und die Entwicklung der KI-Komponente beziehen, auch durch Maßnahmen der klassischen IT-Sicherheit, wie sie in dieser Dimension angebracht werden, abgeschwächt werden sollte. In der Regel ist ein Angriff auf die KI-Anwendung umso erfolgreicher, je genauer der*die Angreifer*in das ML-Modell kennt. Somit hängt die Integrität der KI-Anwendung indirekt mit der Vertraulichkeit von Daten (Gewichten, Trainingsdaten, etc.) zusammen, und Maßnahmen zum Datenschutz tragen entsprechend zur Abschwächung von Sicherheitsrisiken bei. Um eine Dopplung von Maßnahmen in der **Dimension: Sicherheit (SI)** und in der **Dimension: Datenschutz (DS)** zu vermeiden, werden relevante klassische Maßnahmen zur Vertraulichkeit, wie etwa Verschlüsselung, im vorliegenden Risikogebiet angeführt. Die **Dimension: Datenschutz (DS)** hingegen fokussiert auf KI-spezifische Maßnahmen zur Vertraulichkeit, die sich auf das Design und die Modellbildung des in der KI-Anwendung implementierten Lernverfahrens beziehen.

Zum anderen kann die Integrität einer KI-Anwendung in dem Ausmaß verletzt werden, dass ihre Funktionalität (dauerhaft) verändert wird. Dieses Szenario kann beispielsweise eintreten, wenn Angreifer*innen unautorisierte Änderungen am Code oder den Gewichten vornehmen. Anders als bei klassischen IT-Systemen ist es bei KI-Anwendungen jedoch unter Umständen auch möglich, Funktionsänderungen durch gezielte Manipulation der Datenbasis, sogenanntes *Data Poisoning*, herbeizuführen. Bei KI-Anwendungen, die ihre Entscheidungsregeln/Modelle etwa online auf Nutzereingaben (weiter-)lernen, kann dies bereits durch gezieltes An- bzw. Abfragen der Anwendung erreicht werden.

Verfügbarkeit im Kontext der Informationssicherheit bedeutet, dass die KI-Anwendung zeitnah und wie vorgesehen ausgeführt wird bzw. abrufbar ist. Bei klassischen IT-Systemen liegt Nichtverfügbarkeit in der Regel entweder dann vor, wenn die Hardware aufgrund einer hohen Anzahl an Anfragen überlastet ist, oder wenn sich das System in einem Fehlermodus befindet und Anfragen dementsprechend überhaupt nicht bearbeitet. Zwar ist ersteres Szenario (im Sinne von *Denial-of-Service-Attacks*) für KI-Anwendungen relevant, da sie in der Regel rechenintensiv sind und sich die Anzahl an Nutzeranfragen oftmals, z. B. bei KI-Anwendungen mit öffentlicher Schnittstelle wie etwa einem Online-Übersetzungsdienst, stark verändern kann. Die Skalierbarkeit sollte somit in der Architektur einer KI-Anwendung berücksichtigt werden, jedoch liegt die Hardware nicht im Fokus dieses Prüfkatalogs. Das zweite Szenario, d. h. Nichtverfügbarkeit aufgrund eines Fehlermodus sowie andere Arten von Ausfällen werden, sofern sie zum Prüfgegenstand des Prüfkatalogs gehören, in der **Dimension: Verlässlichkeit (VE)** und im **Risikogebiet: Funktionale Sicherheit (FS)** adressiert. Einschränkungen der Verfügbarkeit, die durch Fehler bzw. Ausfälle der Einbettung bei korrekter Funktion der KI-Komponente verursacht werden, werden in diesem Prüfkatalog nicht betrachtet.

Zusätzlich zu den Szenarien der Nichtverfügbarkeit bei klassischen IT-Systemen besteht bei KI-Anwendungen eine weitere KI-spezifische Form der Nichtverfügbarkeit. Sie rührt daher, dass die KI-Komponente ggf. ihre Entscheidungsregeln/das Modell während des Betriebs (weiter-)lernt und dementsprechend ihre Funktion verändern kann. Führt das Weiterlernen dazu, dass die KI-Anwendung ihre ursprüngliche Form und Qualität im Laufe des Betriebs verliert, so stellt dies eine Form der Nichtverfügbarkeit dar. Dies lässt sich am Beispiel eines Chatbots[79] verdeutlichen, der online durch Nutzerinteraktion (weiter-)lernt. Wird er massenhaft mit ungewünschten Eingaben wie etwa Kraftausdrücken konfrontiert, sodass er selbst überwiegend unangemessene Ausdrücke ausgibt, dann ist dies – auch wenn der Chatbot weiterhin zeitnah Nachrichten beantwortet und sich nicht erkennbar in einem Fehlermodus befindet – als Beeinträchtigung seiner Verfügbarkeit aufzufassen, da seine KI-basierte Funktionalität in ihrer ursprünglichen Form und Qualität nicht mehr vorhanden ist. Dieses Beispiel von *Data Poisoning* verdeutlicht zudem den Zusammenhang zwischen Risiken hinsichtlich der Integrität und Verfügbarkeit einer KI-Anwendung.

Ähnlich wie im Fall adversarialer Beispiele, ist die Veränderung der KI-Anwendung während des Betriebs (im Sinne von *Model Drift*) in ihrer Ursache der **Dimension: Verlässlichkeit (VE)** zuzuordnen. Da eine starke Funktionsänderung jedoch auch ein Sicherheitsrisiko darstellt, insbesondere wenn sie durch gezielte Attacken wie etwa *Data Poisoning* herbeigeführt wird, wird ein *Drift* der KI-Anwendung auch in diesem Risikogebiet behandelt. Die Maßnahmen zur Verfügbarkeit in diesem Risikogebiet beziehen sich jedoch nicht auf das Design und die Entwicklung der KI-Komponente (wie es in der **Dimension: Verlässlichkeit (VE)** der Fall ist), sondern orientieren sich an Maßnahmen der klassischen IT-Sicherheit, um potenzielle Nichtverfügbarkeit einzuschränken und die Verfügbarkeit der KI-Anwendung ggf. schnell wiederherzustellen. Insbesondere liegt der Fokus auf (potenziell finanziellen) Schäden, die dem*der Betreiber*in der KI-Anwendung aufgrund von Nichtverfügbarkeit entstehen können. Maßnahmen, die darauf abzielen, Schäden an Nutzer*innen oder der Umwelt während oder aufgrund von Nichtverfügbarkeit im Sinne eines Ausfalls der KI-Anwendung zu verhindern bzw. abschwächen, werden hingegen im **Risikogebiet: Abfangen von Fehlern auf Modellebene (AF)** der Dimension Verlässlichkeit sowie im **Risikogebiet: Funktionale Sicherheit (FS)** behandelt. Die zwei genannten Risikogebiete enthalten unter anderem Maßnahmen, um Schäden während eines Ausfalls systemrelevanter Komponenten, z. B. aufgrund schädlicher Eingaben, zu vermeiden.

8.2.1 Risikoanalyse und Zielvorgaben

[SI-R-IV-RI-01] Risikoanalyse und Zielvorgaben

Anforderung: Do

- **Risikoanalyse Integrität:** Für die zu betrachtende KI-Anwendung werden Gefährdungen der Integrität spezifiziert. Typische Gefährdungen der Integrität sind solche, die zu unerwünschter Modifikation der KI-Anwendung und insbesondere ihrer Ausgaben führen. Angesichts des vorliegenden Anwendungskontextes wird untersucht, welche Risiken in Bezug auf die Manipulation von Trainingsdaten oder der Trainingsumgebung, sowie die Manipulation des ML-Modells bestehen. Außerdem werden plausible Angriffsmöglichkeiten auf die KI-Anwendung analysiert und deren Eintrittswahrscheinlichkeit abgeschätzt. Hierbei sollte auch das Risiko eines Datenabflusses bezüglich Trainingsdaten oder Modellparametern berücksichtigt werden, sofern sich hieraus ein erhöhtes Manipulationsrisiko der KI-Anwendung ergibt. Zuletzt wird untersucht, welche Auswirkungen die Verletzung der Integrität der KI-Anwendung haben kann und insbesondere, welche Schäden daraus resultieren können.
- **Risikoanalyse Verfügbarkeit:** Für die zu betrachtende KI-Anwendung werden Gefährdungen der Verfügbarkeit spezifiziert. Angesichts des vorliegenden Anwendungskontexts werden Risiken in Bezug auf Angriffe, die eine Funktionsänderung bewirken können, in Bezug auf Verzögerungen oder Unterbrechungen in der Kommunikation zwischen KI-Komponente und Einbettung (etwa bei Einbindung von Online-Diensten) sowie in Bezug auf die Skalierbarkeit auf Anfragen untersucht. Dabei ist insbesondere die Nutzungshäufigkeit und

79 Beuth, P. (März 2016). Twitter-Nutzer machen Chatbot zur Rassistin. Zeit Online. https://www.zeit.de/digital/internet/2016-03/microsoft-tay-chatbot-twitter-rassistisch (letzter Aufruf: 16.06.2021)

-dauer der KI-Anwendung zu berücksichtigen. Ferner wird abgeschätzt, welche (nicht physischen) Schäden bei Nichtverfügbarkeit der KI-Anwendung potenziell entstehen können. (Physische Schäden werden im **Risikogebiet: Funktionale Sicherheit (FS)** behandelt.)

- **Zielvorgaben:** Basierend auf der Risikoanalyse werden Zielvorgaben für die Absicherung und Prüfung bezüglich Integrität und Verfügbarkeit dokumentiert. Die Zielvorgaben beschreiben insbesondere, unter welchen Umständen die Beherrschung der in der Risikoanalyse als relevant identifizierten Gefährdungen bzw. Risiken erreicht ist.

8.2.2 Kriterien zur Zielerreichung

Basierend auf den identifizierten Gefährdungen des Risikogebiets Integrität und Verfügbarkeit sollen entsprechende Absicherungsmaßnahmen getroffen werden. Um in der abschließenden Bewertung der Maßnahmen objektiv überprüfen zu können, ob die identifizierten Risiken erfolgreich abgeschwächt wurden, müssen die in **[SI-R-IV-RI-01]** beschriebenen Zielvorgaben in quantitative Kriterien übersetzt werden. Dazu wird unter anderem eine spezifischere Definition der potenziellen Schadenshöhe betrachtet (siehe **[SI-R-IV-KR-01]**), mit der das Restrisiko bewertet wird.

[SI-R-IV-KR-01] Quantifizierung des vertretbaren Risikos

Anforderung: Do

- Es werden Kriterien festgelegt, nach denen die Risiken in Bezug auf Integrität und Verfügbarkeit basierend auf der Höhe potenzieller Schäden und ihrer Eintrittswahrscheinlichkeit beurteilt werden. Bei der Definition der Kriterien sollten mindestens die folgenden Aspekte berücksichtigt werden:

  Schadenshöhe:
  - Relevanz der unerwünscht preisgegebenen bzw. entwendeten Daten/Informationen für die Integrität der KI-Anwendung (im Sinne von Angreifbarkeit)
  - Umfang der Verletzung der Verfügbarkeit (d. h. ist die gesamte KI-Anwendung oder eine Teilfunktion betroffen?)
  - Dauer der Nichtnutzbarkeit (eines Teils) der Anwendung
  - Anzahl der im Fall von Nichtverfügbarkeit betroffenen Personen und abhängigen Systeme
  - Maximal zulässige Latenz (Anforderungen an die Latenz ergeben sich in der Regel aus dem Anwendungskontext)
  - Kosten aufgrund der Nichtnutzbarkeit

  Eintrittswahrscheinlichkeit, unter Berücksichtigung möglicher Ursachen wie:
  - Angriffe
  - Unautorisierter Zugriff auf Modell/Daten
  - Eingeschränkte Skalierbarkeit
  - Hohe Anzahl nicht zweckgemäßer Nutzung (Scherzanfragen, »Trolling«), bspw. durch automatisierte Anfragen oder Schadprogramme

- Unter Verwendung der geschätzten Eintrittswahrscheinlichkeiten wird der erwartete Schaden über die Lebensdauer der KI-Anwendung (bei unbeschränktem Einsatz ein Jahresmittel) ermittelt. Hierbei ist zu berücksichtigen, dass gerade bei automatisierten Verletzungen, etwa durch zur Verfügung stehende Skripte zum Angriff der KI-Anwendung, von einem gehäuften Auftreten auszugehen ist. Neben dem erwarteten mittleren Schaden sollten auch mögliche Abweichungen, etwa anhand von W*orst Case* Betrachtungen, berücksichtigt werden.
- In Bezug auf den erwarteten Schaden werden für die obigen Kriterien Grenzwerte festgelegt, unter deren Einhaltung ein vertretbares Risiko bzw. ein vertretbarer erwarteter Schaden hergestellt ist. Es wird nachvollziehbar argumentiert, dass die festgelegten Kriterien und Grenzwerte für den vorliegenden Anwendungskontext angemessen und ausreichend sind, und dass sie die Zielvorgaben in **[SI-R-IV-RI-01]** abbilden.

[SI-R-IV-KR-02] Zugriffsmöglichkeiten auf Daten

Anforderung: Do

Für alle Daten, die im Zusammenhang mit der KI-Anwendung stehen und über deren Zugriff bzw. Veränderung die Qualität der KI-Anwendung vermindert werden könnte, wird basierend auf den in **[SI-R-IV-RI-01]** ermittelten Risiken festgehalten,

- für welchen Personenkreis,
- in welcher Häufigkeit,
- zu welchem Zweck,
- und unter welchen sonstigen Voraussetzungen.

Einblick in das Datum oder eine Modifikation dessen möglich sein soll bzw. wann dies im Gegenzug nicht erlaubt ist.

[SI-R-IV-KR-03] Anzahl an An-/Abfragen der KI-Anwendung

Anforderung: Do

- Unter Umständen können Risiken hinsichtlich gezielter Attacken, Funktionsänderung etwa durch *Data Poisoning*, sowie *Denial-of-Service-Attacks* eine Beschränkung der Anzahl an Nutzeranfragen erforderlich machen. Ist dies der Fall, so sind die An-/Abfragemöglichkeiten der KI-Anwendung in Art und Umfang (pro Nutzer*in oder insgesamt) derart gestaltet, dass Integrität und Verfügbarkeit gewährleistet sind. Andernfalls ist zu begründen, weshalb eine Beschränkung der Anfragen von Nutzer*innen an die KI-Anwendung nicht als erforderlich erachtet wird.

8.2.3 Maßnahmen

[SI-R-IV-MA-01] Sicherheitsrichtlinien und Nutzungsanweisungen

Anforderungen: Do | Pr

- Es liegt eine Dokumentation vor, in der Sicherheitsziele bzgl. der Integrität und Verfügbarkeit aus den Unternehmenszielen, Geschäftsprozessen, relevanten Gesetzen, Verordnungen und möglichen Gefährdungen basierend auf der Risikoanalyse **[SI-R-IV-RI-01]** zusammengefasst werden. Diese Sicherheitsrichtlinie enthält außerdem strategische Vorgaben, wie diese Ziele erreicht werden sollen.
- Es ist ein Prozess etabliert und dokumentiert, der Nutzer*innen und Entwickler*innen der für Risiken bezüglich Integrität und Verfügbarkeit sensibilisiert und auf den korrekten Umgang in Bezug auf die Integrität von Daten, Modell und Einbettung sowie in Bezug auf die Verfügbarkeit der KI-Anwendung hinweist.
- Es sind Maßnahmen beschrieben, die dafür sorgen, dass alle Nutzer*innen diese zur Kenntnis nehmen. Ferner wird dargelegt, inwiefern diese Maßnahmen zur Abschwächung bzw. Beherrschung der identifizierten Risiken gemäß **[SI-R-IV-KR-01]** beitragen.
  (angelehnt an BSI-C5 SA-01)

8.2.3.1 Daten

[SI-R-IV-MA-02] Datenintegrität

Anforderung: Do

- Es liegt eine Dokumentation darüber vor, welche Maßnahmen, wie etwa Signaturen oder Prüfsummen, ergriffen werden, um die Integrität der Trainingsdaten, des trainierten Modells (d. h. Hyperparameter und Gewichte) und weiterer gespeicherter bzw. im Betrieb neu einkommender Daten in der Trainings- und Produktivumgebung zu gewährleisten.

Maßnahmen zur Gewährleistung der Datenintegrität können unter anderem verhindern, dass Angreifer*innen Gewichte oder Datensätze manipulieren, um die KI-Anwendung für bestimmte Arten von Angriffen empfänglich zu machen (*Data Poisoning*).

[SI-R-IV-MA-03] Vertraulichkeit der Daten

Anforderung: Do

- Es liegt eine Dokumentation darüber vor, welche Maßnahmen, wie etwa Verschlüsselung, ergriffen werden, um die Vertraulichkeit der Trainingsdaten, des trainierten Modells (d. h. Hyperparameter und Gewichte) und weiterer gespeicherter bzw. im Betrieb neu einkommender Daten in der Trainings- und Produktivumgebung zu gewährleisten.
  Maßnahmen zur Vertraulichkeit des Modells erschweren es Angreifer*innen beispielsweise, spezifische Attacken (sog. *White-Box*-Attacken) zu konstruieren, die etwa eine bestimmte Ausgabe erzwingen könnten.

[SI-R-IV-MA-04] Sicherung und Wiederherstellung von Daten

Anforderung: Do

- Es liegt eine Dokumentation von technischen und organisatorischen Maßnahmen zur Vermeidung von Datenverlusten bzgl. Trainingsdaten, trainiertem Modell oder sonstigen Einstellungen und Daten vor, damit diese nach einem verlustbedingten Ausfall in angemessener Zeit wieder zur Verfügung stehen. Insbesondere bei ML-Modellen, die inkrementell lernen, sollte jederzeit ein *Roll-Back* zur letzten Version des Modells möglich sein. Die ergriffenen Maßnahmen zur regelmäßigen Sicherung und Wiederherstellung von Daten, sowie deren Umfang, Dauer und Häufigkeit sind dokumentiert. Falls dies bereits an anderer Stelle dokumentiert ist, z. B. in der **Dimension: Transparenz (TR)**, so ist auf die entsprechende Stelle zu verweisen.

8.2.3.2 KI-Komponente

Für das Risikogebiet Integrität und Verfügbarkeit sind keine risikomindernden Maßnahmen, die sich auf Entwicklung und Modellbildung der KI-Komponente beziehen, vorgesehen. Jedoch steht, wie in der Einleitung beschrieben, die Integrität der KI-Anwendung indirekt in Zusammenhang mit der Vertraulichkeit von Daten. Maßnahmen zum Schutz von Daten, die sich auf die KI-Komponente beziehen, werden in der **Dimension: Datenschutz (DS)** beschrieben. Hinzu kommt bei weiterlernenden KI-Anwendungen die Gefahr von *Model Drift*. Maßnahmen auf Ebene der KI-Komponente finden sich hierzu im **Risikogebiet: Beherrschung der Dynamik (BD)** der Dimension Verlässlichkeit.

8.2.3.3 Einbettung

[SI-R-IV-MA-05] Physischer Schutz des Speicherorts

Anforderung: Do

- Es liegt eine Dokumentation darüber vor, welche Orte zum Speichern von im Zusammenhang mit der KI-Anwendung stehenden Daten genutzt werden. Insbesondere sind die Speicherorte von Trainingsdaten, Gewichten und Hyperparametern des trainierten Modells, sowie weiteren Daten wie z. B. Protokollierungsdaten zu beschreiben.
- Darüber hinaus liegt eine Dokumentation vor, welche Maßnahmen ergriffen werden, um die Daten, sowie die Trainings- und Produktivumgebung vor unberechtigtem physischen Zutritt und einhergehendem Diebstahl oder Schaden angemessen zu schützen. Im Fall externer Anbieter, z. B. Clouddienste, wird dargelegt, dass diese ein hinreichendes Sicherheitskonzept gewährleisten.
  (basierend auf BSI-C5 PS)

[SI-R-IV-MA-06] Schutz vor Schadprogrammen

Anforderungen: Do | Pr

- Es liegt eine Dokumentation vor, die beschreibt, welcher Schutz vor Schadprogrammen in der Trainings- und Produktivumgebung vorhanden ist. Hierbei sind auch KI-spezifische Schadprogramme, wie Programme zur Ausführung von adversarialen Attacken (siehe **Dimension: Verlässlichkeit (VE)**), zu berücksichtigen.
- Ferner ist ein Prozess zur Beobachtung und Identifizierung möglicher neuartiger Schadprogramme, sowie zur Überprüfung des State of the Arts der Schutzsoftware inklusive regelmäßiger Updates etabliert.

[SI-R-IV-MA-07] Kommunikationssicherheit

Anforderung: Do

- Es liegt eine Dokumentation darüber vor, durch welche Maßnahmen eine sichere, vertrauliche Kommunikation hergestellt wird. Dabei ist sowohl auf die Kommunikation des*der Nutzer*in mit der KI-Anwendung, sowie auf die Kommunikation innerhalb der Anwendung, z. B. zwischen KI-Komponente und Einbettung, einzugehen. Je nach Schutzbedarf sind bestehende Normen/Standards der Informationssicherheit heranzuziehen, die entsprechende Maßnahmen aufzeigen.
  (basierend auf BSI-C5 KOS-01)

[SI-R-IV-MA-08] Timeout der KI-Anwendung

Anforderung: Do

- Es liegt eine Dokumentation darüber vor, dass eine Zeitüberschreitung bzw. ein Timeout innerhalb der KI-Komponente oder in der Interaktion der KI-Komponente mit der Einbettung kein unvertretbares Risiko darstellt. Zum einen ist dokumentiert, inwiefern gewährleistet werden kann, dass die KI-Komponente immer in der erforderlichen Zeit antwortet. Ist dies nicht der Fall, z. B. weil Ressourcen nicht dauerhaft vorgehalten werden können, die Antwortzeit nicht deterministisch ist, oder es Verzögerungen in der Kommunikation der KI-Komponente zur Einbettung gibt, so ist darzulegen, wie mit einer Zeitüberschreitung umgegangen wird, ohne dass Risiken unvertretbar werden.
- Zum anderen wird für KI-Anwendungen, in denen direkte Feedbackschleifen zwischen klassischen Softwarekomponenten der Einbettung und der KI-Komponente implementiert sind, untersucht, inwiefern Feedbackschleifen mit anderen Komponenten der Einbettung in der erforderlichen Zeit verlassen/beendet werden.

[SI-R-IV-MA-09] Testen der Skalierbarkeit

Anforderungen: Do | Te

- Es liegt eine Dokumentation vor, die festhält, wie die KI-Anwendung mit einer Fülle an Anfragen umgeht, bei denen ein hohes Risiko der Nichtverfügbarkeit besteht. Ein typisches Beispiel, bei dem ein Risiko der Nichtverfügbarkeit besteht, sind ausgelagerte KI-Komponenten, etwa Smart-Home-Systeme, die bei der Bearbeitung von Nutzeranfragen auf serverseitige Rechenkapazitäten zurückgreifen. Aber auch Chat-Bot-Systeme, z. B. im Bereich der automatisierten Akquise oder Kundenbetreuung, auf Webseiten können von stark geändertem Nutzerverhalten beeinträchtigt werden, etwa wenn die Nachfrage oder das Interesse nach einem Produkt unerwartet stark steigt. Für KI-Anwendungen gibt es, im Vergleich zu Konzepten wie Software-as-a-Service, daher häufig höhere Ansprüche an die »Echtzeitverfügbarkeit« (siehe vorheriges Beispiel) bei gleichzeitig hoher Rechenlast, da Informationen vor der Bereitstellung verarbeitet werden müssen.
- Es werden Tests durchgeführt und dokumentiert, die gezielt einen Ausfall der KI-Anwendung provozieren. Die untersuchten Testszenarien und ggf. dazu verwendeten Datensätze sind zu dokumentieren. Es wird begründet, dass die Testergebnisse ein angemessenes Verhalten in Bezug auf die Skalierbarkeit nach den definierten Zielen in **[SI-R-IV-KR-01]** attestieren.

8.2.3.4 Maßnahmen für den Betrieb

[SI-R-IV-MA-10] Identitäts- und Berechtigungsmanagement

Anforderungen: Do | Pr

- Es liegt eine Dokumentation darüber vor, dass ein Rollen- und Rechtekonzept, sowie ein Prozess zur Verwaltung von Zugangs- und Zugriffsberechtigungen für die Trainings- und Produktivumgebung definiert sind. Die folgenden Bereiche werden dabei adressiert:
  - Die Vergabe und Änderung von Zugriffsberechtigungen für Daten, Trainingsumgebung, trainiertes Modell und Einbettung auf Basis des Prinzips der geringsten Berechtigung und wie es für die Aufgabenwahrnehmung notwendig ist
  - Die Registrierung der Nutzer*innen und Maßnahmen, die eine eindeutige Benutzererkennung sicherstellen
  - Sowie unter Umständen (siehe **[SI-R-IV-KR-03]**) die Beschränkung der Anzahl an Ab- bzw. Anfragen, die von Nutzer*innen an die KI-Anwendung gestellt werden können. Art und Umfang der Beschränkung sind zu beschreiben. Insbesondere ist darzulegen, dass sich die Beschränkung quantitativ am Risiko des Integritätsverlustes orientiert, d. h. an der Anzahl an Ab-/Anfragen, die erforderlich sind, um eine sensitive Information zu rekonstruieren oder die Datenbasis derart zu manipulieren, dass die Qualität der KI-Anwendung sinkt
    (angelehnt an BSI-C5 IDM-01)
- Es wird dargelegt, inwiefern durch die hier getroffenen Maßnahmen die Anforderungen **[SI-R-IV-KR-02]** und **[SI-R-IV-KR-03]** erfüllt werden.

[SI-R-IV-MA-11] Protokollierung und Überwachung

Anforderungen: Do | Pr

- Es liegt eine Dokumentation darüber vor, durch welche technischen und organisatorischen Maßnahmen definierte Ereignisse in der Trainings- und Produktivumgebung, die die Integrität oder Verfügbarkeit der KI-Anwendung beeinträchtigen können, protokolliert und überwacht werden. Folgende Punkte sollten mindestens betrachtet werden:
  - Die Aktivierung, das Stoppen und Pausieren der Protokollierungen
  - Das Erstellen, Ändern oder Löschen von Benutzern bzw. Benutzerberechtigungen bzgl. der in **[SI-R-IV-MA-10]** definierten Bereiche
  - Das Erstellen oder Einfügen von unter **[SI-R-IV-KR-01]** als relevant erachteter Daten
  - Das Trainieren des Modells und Erstellen einer aktuellen Version des Modells
  - Sofern dies nicht im Widerspruch zur **Dimension: Datenschutz (DS)** steht, die Nutzeranfragen und die zugehörige Ausgabe der KI-Anwendung, ggf. ohne Bezug auf den*die Nutzer*in (es sollte überprüft werden, ob durch die Art der Anfrage Rückschlüsse auf den*die Nutzer*in gezogen werden können und ob dies kritisch wäre). Die kann beispielsweise im Kontext der Bearbeitung von Sicherheitsvorfällen, der Anpassung der KI-Anwendung an die allgemeine Nutzung und der Nachvollziehbarkeit von Entscheidungen der KI-Anwendung hilfreich bzw. relevant sein.
  - Metadaten der Nutzer*innen, getrennt von deren Anfragen. Diese können etwa zur Behebung von Störungen und Fehlern, sowie zum Bearbeiten von Sicherheitsvorfällen hilfreich bzw. relevant sein. Es wird beschrieben, inwiefern eine kommerzielle Nutzung der Metadaten stattfindet und zu welchem Zeitpunkt diese gelöscht werden.

  Falls die Überwachung bzw. Protokollierung der gelisteten Ereignisse bereits an anderer Stelle dokumentiert ist (z. B. in der **Dimension: Transparenz (TR)** oder in der **Dimension: Verlässlichkeit (VE)**), so kann stattdessen auch der entsprechende Verweis eingefügt werden.
- Außerdem werden Verletzungen der Integrität oder Verfügbarkeit protokolliert. Zum einen werden Art, Umfang und, falls möglich, Ursache des Vorfalls dokumentiert. Zum anderen wird festgehalten, wie die KI-Anwendung mit dem Vorfall umgegangen ist.

- Die Protokolle werden bei unerwarteten oder auffälligen Ereignissen durch autorisiertes Personal anlassbezogen überprüft, um eine zeitnahe Untersuchung von Störungen und Sicherheitsvorfällen sowie das Einleiten geeigneter Maßnahmen zu ermöglichen. Personenbezogene Daten werden nach den geltenden datenschutzrechtlichen Anforderungen (siehe auch **Dimension: Datenschutz (DS)**) gespeichert und geschützt. (angelehnt an BSI-C5 RB-10 und 11)

[SI-R-IV-MA-12] Organisation der Informationssicherheit

Anforderung: Pr

- Es existiert ein Prozess zur Organisation der Informationssicherheit. Der Betreiber initiiert, steuert und überwacht ein Managementsystem zur Informationssicherheit, das die Verfügbarkeit der KI-Anwendung beachtet, angelehnt an den Standard ISO-27001 zur Planung, Umsetzung, Aufrechterhaltung und kontinuierlichen Verbesserung eines Rahmenwerks zur Informationssicherheit der Anwendung.

[SI-R-IV-MA-13] Vorgehen bei Verlust von Integrität oder Verfügbarkeit

Anforderungen: Do | Pr

- Es wird dokumentiert, welche Maßnahmen ergriffen werden, um im Fall einer Verletzung von Integrität oder Verfügbarkeit potenzielle weitere Schäden zu vermeiden bzw. zu minimieren.
- Es werden die Prozesse beschrieben, die beim festgestellten Verlust greifen. Im Falle mangelnder Integrität kann dies, je nach Schwere, etwa eine diagnostische Prüfung der KI-Anwendung, die Abschaltung oder die Überführung in einen *Fail-Safe State* beinhalten. Falls beispielsweise die Gewichte des ML-Modells offengelegt wurden, kann das Risiko einer *White-Box*-Attacke etwa durch Neutraining oder *Roll-Back* auf eine vorige Version abgeschwächt werden.

[SI-R-IV-MA-14] Wiederherstellung der KI-Komponente

Anforderung: Pr

- Es existiert ein Prozess dazu, die KI-Komponente bei Nicht-Funktionsfähigkeit oder unerwünschten Änderungen innerhalb der Einbettung neu aufzusetzen, bei Bedarf unter Nutzung der Datensicherungen (insbesondere der gespeicherten Gewichte und der Architektur) aus **[SI-R-IV-MA-04]**. Hierbei kann auf Maßnahmen der klassischen IT-Sicherheit verwiesen werden.

[SI-R-IV-MA-15] Erkennen des Verlusts von Integrität oder Verfügbarkeit

Anforderungen: Do | Pr

- Es liegt eine Dokumentation von Maßnahmen zum Erkennen des Verlusts von Integrität oder Verfügbarkeit vor. Unter anderem kann dabei auf Maßnahmen zum Erkennen von *Model Drift* verwiesen werden (siehe **Risikogebiet: Beherrschung der Dynamik (BD)** der Dimension Verlässlichkeit). Außerdem sind ggf. Meldemöglichkeiten seitens der Nutzer*innen und Prozesse zur zuverlässigen, angemessenen und zeitnahen Behandlung von Nutzer*innen-Anfragen zu beschreiben.

8.2.4 Gesamtbewertung

[SI-R-IV-BW] Gesamtbewertung

Anforderung: Do

- Es wird ausführlich begründet, dass die Kriterien **[SI-R-IV-KR-02]** und **[SI-R-IV-KR-03]** erfüllt sind. Außerdem wird dargelegt, dass das Restrisiko gemäß den Kriterien **[SI-R-IV-KR-01]** vertretbar ist.
- Sofern nicht alle in **[SI-R-IV-KR-01]** bis **[SI-R-IV-KR-03]** spezifizierten Anforderungen erfüllt werden, sind die Abweichungen zu dokumentieren. Dies gilt ebenfalls für nur teilerfüllte Anforderungen, bei denen etwa die Kriterien nicht, oder nicht immer, erreicht wurden.

8.3 Risikogebiet: Beherrschung der Dynamik (BD)

Das Risikogebiet Beherrschung der Dynamik soll sicherstellen, dass die Sicherheit der KI-Anwendung auch im Laufe des Betriebs gewährleistet ist.

Ein maßgebliches Risiko in diesem Kontext ist der *Concept Drift*, der dazu führen kann, dass implementierte Sicherheitsmaßnahmen aufgrund sich im Laufe des Betriebs ändernder Anforderungen oder Umstände nicht mehr greifen oder nicht mehr ausreichend sind. Die Forschung im Bereich der Künstlichen Intelligenz könnte etwa neue Erkenntnisse hervorbringen, die Angreifer*innen zur Entwicklung neuer Schadprogramme und Angriffsmethoden verhelfen. Außerdem können sich durch Änderung äußerer Umstände im Laufe des Betriebs neue Gefährdungen der Sicherheit entwickeln. Nicht zuletzt können durch gesetzliche oder regulatorische Änderungen neue Anforderungen an die Sicherheit gestellt werden.

Ein Beispiel für ein Risiko durch *Concept Drift* stellt ein Dienst zur automatisierten Übersetzung digitaler Kommunikation dar, der durch verändertes Nutzerverhalten eine stark erhöhte Nachfrage erfährt und die Verfügbarkeit nicht mehr gewährleisten kann (vgl. **[SI-R-IV-MA-09]**). Andere Beispiele ergeben sich durch den Austausch von Aktuatoren eines cyber-physikalischen Systems, die Parameter mit anderen Toleranzen erwarten, sodass hier Schwellwerte für die Mitigation (vgl. **[SI-R-FS-MA-08]**) angepasst werden müssen.

Ein weiterer Aspekt der Dynamik von KI-Anwendungen ist der *Model Drift*. Dieser spielt für dieses Risikogebiet jedoch eine untergeordnete Rolle, da Herausforderungen durch das Weiterlernen des Modells auf neu aufgenommen Daten bereits im **Risikogebiet: Beherrschung der Dynamik (BD)** in der Dimension Verlässlichkeit umfassend behandelt werden. Zudem werden im **Risikogebiet: Abfangen von Fehlern auf Modellebene (AF)** und im **Risikogebiet: Funktionale Sicherheit (FS)** Maßnahmen angeführt, die zum Abfangen KI-bezogener sicherheitsrelevanter Fehler bzw. Störungen beitragen, zu denen insbesondere die Verringerung der Performanz während des Betriebs zählt. Im vorliegenden Risikogebiet werden deshalb nur solche Maßnahmen bezüglich *Model Drift* behandelt, die sich auf Sicherheitsmechanismen beziehen, die von der Leistungsfähigkeit der KI-Anwendung abhängen. Insbesondere sind diese Sicherheitsmechanismen bei schwächerer Performanz der KI-Komponente entsprechend zu schärfen.

8.3.1 Risikoanalyse und Zielvorgaben

In der Risikoanalyse werden Gefährdungen des Risikogebiets Beherrschung der Dynamik bezüglich Sicherheit untersucht. Typische Gefährdungen bzgl. der Beherrschung der Dynamik sind solche Sicherheitsgefährdungen, die aufgrund von

- *Concept Drift* durch eine Änderung von Rahmenbedingungen,
- Änderungen im Nutzerverhalten,
- Veränderte Anforderungen durch Änderungen zugrundeliegender Frameworks oder Hardware,

entstehen können.

Diese Gefährdungen können dabei jegliches der zuvor betrachteten Risikogebiete, **Risikogebiet: Funktionale Sicherheit (FS)** sowie **Risikogebiet: Integrität und Verfügbarkeit (IV),** betreffen. Das genaue Vorgehen in der Risikoanalyse für das Risikogebiet Beherrschung der Dynamik wird im folgenden Abschnitt beschrieben.

[SI-R-BD-RI-01] Risikoanalyse und Zielvorgaben

Anforderung: Do

- **Risikoanalyse:** In der Risikoanalyse **[VE-R-BD-RI-01]** in der Dimension Verlässlichkeit wurde bereits das Risiko von *Model* und teilweise auch von *Concept Drift* untersucht und kann an dieser Stelle referenziert werden. Es ist jedoch zu analysieren, in welchem Ausmaß die implementierten Sicherheitsmaßnahmen von der Performanz der KI-Anwendung abhängen und dementsprechend bei *Drift* angepasst werden müssen. Ebenso sind regulatorische Änderung in diese Risikoanalyse miteinzubeziehen.

Ferner ist zu untersuchen, welche äußeren Umstände im vorliegenden Anwendungskontext einen Einfluss auf die Sicherheitsanforderungen der KI-Anwendung haben können. Dabei wird analysiert, welche Angriffsmöglichkeiten auf die KI-Anwendung durch neuartige bzw. absehbare Schadprogramme bestehen. Außerdem wird untersucht, inwiefern Änderungen des Nutzerverhaltens Einfluss auf die Sicherheit haben können. Weitere kontextspezifische Gefährdungsszenarien für die Beherrschung der Dynamik sind ggf. zu ergänzen. Abschließend werden potenzielle Schäden, die sich aus Sicherheitslücken aufgrund von Weiterlernen im Betrieb sowie geänderter Sicherheitsanforderungen aufgrund äußerer Rahmenbedingungen ergeben können, und deren Eintrittswahrscheinlichkeit angesichts des vorliegenden Anwendungskontextes abgeschätzt.

- **Zielvorgaben:** Basierend auf der Risikoanalyse werden Ziele für das Vermeiden, Erkennen, Erfassen und Behandeln potenzieller Sicherheitsvorfälle aufgrund von *Model* oder *Concept Drift* aufgestellt.

8.3.2 Kriterien zur Zielerreichung

Basierend auf den Gefährdungen des Risikogebiets Beherrschung der Dynamik sollen entsprechende Absicherungsmaßnahmen getroffen werden. Um in der abschließenden Bewertung der Maßnahmen objektiv überprüfen zu können, ob die vorhandenen Risiken erfolgreich mitigiert wurden, muss die in **[SI-R-BD-RI-01]** beschriebene Zielsetzung in quantitative Kriterien übersetzt werden.

[SI-R-BD-KR-01] Rahmenbedingungen zum Umgang mit Sicherheitsrisiken unter Dynamik

Anforderung: Do

- Es sind Kriterien zu dokumentieren, nach denen der Umgang mit Sicherheitsrisiken unter Dynamik bewertet wird. Die Kriterien können quantitativer oder qualitativer Natur sein und sollten mindestens umfassen:
  - Prüfintervall zur Überprüfung der Aktualität der implementierten Sicherheitsmaßnahmen
  - Umfang der Überprüfung
  - Anforderungen an das Nutzerverhalten
  - Schwellwert oder qualitatives Kriterium, ab wann (in Abhängigkeit der Metriken aus der **Dimension: Verlässlichkeit (VE)** sowie der Kriterien aus der Dimension Sicherheit) eine Anpassung der Sicherheitsmaßnahmen erforderlich ist
- Zu jedem aufgestellten Kriterium sind Zielwerte (oder qualitative Zieleigenschaften) anzugeben. Es ist nachvollziehbar zu argumentieren, dass diese Zielsetzung mit den Zielvorgaben konform und für den vorliegenden Anwendungskontext angemessen ist.

8.3.3 Maßnahmen

8.3.3.1 Daten

Für diese Kategorie sind keine Maßnahmen vorgesehen.

8.3.3.2 KI-Komponente

Für diese Kategorie sind keine Maßnahmen vorgesehen.

8.3.3.3 Einbettung

Für diese Kategorie sind keine Maßnahmen vorgesehen.

8.3.3.4 Maßnahmen für den Betrieb

[SI-R-BD-MA-01] Ausbildung und Sensibilisierung von Mitarbeitenden

Anforderung: Do

- Es liegt eine Dokumentation darüber vor, wie sichergestellt ist, dass Mitarbeitende und Dienstleistende sich ihrer Verantwortung in Bezug auf Informationssicherheit bewusst sind. Hierfür werden folgende Maßnahmen in Betracht gezogen:
  - Zuverlässigkeitsprüfung der Mitarbeitenden (z. B. durch die Verifikation der Person durch Personalausweis, Verifikation des Lebenslaufs oder auch die Anfrage eines polizeilichen Führungszeugnisses bei sensiblen Positionen)
  - Beschäftigungsvereinbarung mit Verpflichtung der Einhaltung von Gesetzen, Vorschriften und Regelungen
  - Programm zur Sicherheitsausbildung und Sensibilisierung, dazu gehören regelmäßige Unterweisungen und Schulungen zur sicheren Entwicklung und Instandhaltung der Anwendung zum angemessenen Umgang mit Trainings- und Nutzerdaten, zur regelmäßigen Unterrichtung von möglichen – vor allem KI-spezifischen – Angriffen und zum regelmäßigen Training des Verhaltens bei Auftritt sicherheitsrelevanter Ereignisse
  - Disziplinarmaßnahmen bei Verstößen gegen Richtlinien und Anweisungen
    (angelehnt an BSI-C5 HR-01 bis HR-04)

[SI-R-BD-MA-02] Überwachung äußerer Rahmenbedingungen

Anforderung: Pr

- Es existiert ein Prozess zur Überwachung äußerer Umstände sowie potenzieller neuartiger Gefährdungen der Sicherheit, die eine Anpassung der Sicherheitsanforderungen und -maßnahmen erforderlich machen könnten. Der Prozess umfasst mindestens die Auseinandersetzung mit
  - Fortschritten im Bereich der Forschung, insbesondere der Möglichkeit von Angriffen und Schadprogrammen für die KI-Anwendung,
  - dem Zutreffen der Grundannahmen, die an den vorliegenden Anwendungskontext getroffen werden und nach denen sich die Funktionsweise sowie Sicherheitsmaßnahmen ausrichten,
  - Änderungen des verwendeten Software-Frameworks, sodass z. B. die aktuelle Version des Codes nicht gelesen werden kann, oder Sicherheitslücken entstehen,
  - Änderungen gesetzlicher und regulatorischer Rahmenbedingungen, hierunter Änderungen, die ebenso die **Dimension: Datenschutz (DS)** und **Dimension: Verlässlichkeit (VE)** betreffen können,

  Art und Umfang des Prozesses werden dokumentiert.
- Außerdem werden Reaktionen bzw. nächste Schritte definiert, die bei Feststellung einer relevanten Änderung äußerer Rahmenbedingungen eingeleitet werden.

[SI-R-BD-MA-03] Notfall-Management

Anforderung: Pr

- Es existiert ein Prozess zum Notfall-Management. Ein Notfall kann ein Unfall, Ausfall oder Sicherheitsfall sein. In diesem Prozess werden Notfallkonzepte (vgl. auch **[SI-R-FS-MA-14]**) geplant, implementiert, getestet, überwacht und regelmäßig überprüft und verbessert. Dazu gehört die Erkennung des Notfalls, dessen Behandlung und die Nutzung des Vorfalls zur Verbesserung der Sicherheit. Nach einem Notfall werden durchgeführt:
  - Eine Notfall-Analyse (*Incident Management*): Der Verlauf von eingetretenen Notfällen ist protokolliert und die resultierenden Daten werden genutzt, um die Sicherheit der Anwendung zu verbessern.
  - Eine Überprüfung aller Sicherheitsmaßnahmen auf deren Wirksamkeit.
- Falls es thematische Überschneidungen gibt, so ist auf Konformität mit den Sicherheitsrichtlinien **[SI-R-FS-MA-01]** und **[SI-R-IV-MA-01]** zu achten.
  (angelehnt an BSI-C5 BCM, SIM)

8.3.4 Gesamtbewertung

[SI-R-BD-BW] Gesamtbewertung

Anforderung: Do

- Unter Würdigung der ergriffenen Maßnahmen ist darzulegen, dass Prozesse bzw. Rahmenbedingungen geschaffen wurden, die die Kriterien **[SI-R-BD-KR-01]** erfüllen und somit ein vertretbares Risiko bezüglich der Dynamik der KI-Anwendung in Hinblick auf Sicherheit herstellen.
- Sofern nicht alle in **[SI-R-BD-KR-01]** spezifizierten Anforderungen erfüllt werden, sind die Abweichungen zu dokumentieren. Dies gilt ebenfalls für nur teilerfüllte Anforderungen, bei denen etwa die Kriterien nicht, oder nicht immer, erreicht wurden.

Zusammenfassende Betrachtung

[SI-Z] Zusammenfassende Betrachtung der Dimension

Anforderung: Do

- Falls für diese Dimension ein mittlerer oder hoher Schutzbedarf besteht, ist eine Dokumentation über die verbleibenden Restrisiken zu erstellen. Zunächst werden die Restrisiken aus den verschiedenen Risikogebieten dieser Dimension zusammengefasst. Anschließend wird unter Berücksichtigung des Schutzbedarfs analysiert, ob die identifizierten Restrisiken insgesamt als vernachlässigbar, nicht vernachlässigbar (aber vertretbar) oder unvertretbar zu bewerten sind. Bei dieser Analyse sollten insbesondere die Auswirkungen von Maßnahmen aus der Dimensionen Verlässlichkeit und Datenschutz berücksichtigt werden, falls diese dazu beitragen, Sicherheitsrisiken abzuschwächen oder zu beherrschen. Das Ergebnis der Analyse ist zu erläutern.
- Falls potenziell negative Auswirkungen von Risiken oder Maßnahmen dieser Dimension auf andere Dimensionen, beispielsweise Autonomie und Kontrolle, Transparenz oder Datenschutz, festgestellt wurden, sind diese zu dokumentieren.
- Es wird ein Fazit über die Dimension gezogen, welches insbesondere die Bewertung der Restrisiken enthält.

9. Dimension: Datenschutz (DS)

Beschreibung und Zielsetzung

Aufgrund ihrer Eigenschaften ist es möglich, dass KI-Anwendungen in eine Vielzahl von Rechtspositionen eingreifen. Besonders häufig handelt es sich dabei um Eingriffe in die Privatsphäre bzw. das Recht auf informationelle Selbstbestimmung. So verarbeiten KI-Anwendungen oftmals sensible Informationen, wie zum Beispiel personenbezogene oder persönliche Daten wie Stimmaufnahmen, Fotos oder Videos. Daher ist sicherzustellen, dass die einschlägigen datenschutzrechtlichen Bestimmungen wie etwa die Datenschutz-Grundverordnung (DSGVO) und das Bundesdatenschutzgesetz (BDSG) eingehalten werden. Jedoch können KI-Anwendungen nicht nur ein Risiko für die Privatsphäre des Einzelnen darstellen. Auch (Geschäfts-)Geheimnisse oder lizenzgebundene Daten können betroffen sein, die keine personenbezogenen Daten im Sinne der DSGVO darstellen, aber dennoch u. U. rechtlich schutzwürdig sind. Beispielsweise können Maschinendaten – völlig unabhängig von der Frage, welche Person als Maschinenbediener*in tätig war – Informationen über die Prozessauslastung oder Fehlerquoten beinhalten und damit ein sensibles geschäftsbezogenes Datum darstellen.[80]

Die Herausforderungen im Zusammenhang mit dem Datenschutz sind bei KI-Anwendungen potenziell deutlich größer als bei klassischen IT-Systemen. Dies liegt insbesondere darin begründet, dass KI-Anwendungen oft Daten zusammenführen, die bislang nicht verknüpft waren, und erst durch Maschinelles Lernen neue Methoden der Verknüpfung von Daten erstellen. Je mehr Daten verknüpft werden (*Data Linkage*), umso größer wird das Risiko, Personen oder z. B. konkrete Betriebsstätten auch ohne direkte Angabe entsprechender Attribute identifizieren zu können. So ist es zum Beispiel mit ca. 95 Prozent Verlässlichkeit möglich, Personen an der Art und Weise, wie sie eine Computertastatur bedienen, zu re-identifzieren[81, 82]. Gäbe es nun eine öffentliche (oder käufliche) Datenbank mit der Zuordnung von Tastatur-Anschlagsmustern zu Personen, so wird das Tastatur-Anschlagsmuster zu einem sogenannten *Quasi-Identifier*, der etwa einen Personenbezug ermöglicht.

Mit KI-Methoden können potenziell Personen- oder Geschäftsbezüge bei der Verarbeitung von Text, Sprach- und Bilddaten, sowie protokollierten Nutzungsdaten erstellt werden. Zusätzlich zu den gespeicherten oder verarbeiteten Daten kann jedoch auch das in der KI-Anwendung implementierte ML-Modell selbst ausgespäht werden. Die gezielte systematische Abfrage der KI-Anwendung, um Modellparameter oder andere Modelleigenschaften zu rekonstruieren, bezeichnet man als Modellextraktion. Hat ein*e Angreifer*in die Modellparameter erlangt und würde zusätzlich den Lernalgorithmus und die Struktur des Modells kennen, so könnte die Person versuchen, das Modell nachzubauen. Dies könnte beispielsweise die Wettbewerbssituation eines betroffenen Unternehmens deutlich verschlechtern. Darüber hinaus könnten gezielte Angriffe erstellt werden, um etwa personenbezogene (Trainings-)Daten aus einem Modell zu extrahieren. Beispielsweise ist es möglich, basierend auf

---

80 Die Darstellung in diesem Abschnitt sowie teils in den folgenden Abschnitten ist stark angelehnt an das Kapitel »3.6 Datenschutz« des Whitepapers: Poretschkin, M., et al. (2019). Vertrauenswürdiger Einsatz von Künstlicher Intelligenz. Sankt Augustin: Fraunhofer-Institut für Intelligente Analyse und Informationssysteme IAIS. https://www.ki.nrw/wp-content/uploads/2020/03/Whitepaper_KI-Zertifizierung.pdf (letzter Aufruf: 18.06.2021)

81 Araujo, L. C. F., Sucupira, L. H. R., Lizarraga, M. G., Ling, L. L., & Yabu-Uti, J. B. T. (2005). User authentication through typing biometrics features. IEEE Transactions on Signal Processing, 53(2), 851–855. https://doi.org/10.1109/TSP.2004.839903 (letzter Aufruf: 28.06.2021)

82 Killourhy, K.S., & Maxion, R. (2009). Comparing anomaly-detection algorithms for keystroke dynamics. 2009 IEEE/IFIP International Conference on Dependable Systems & Networks, 125-134. https://doi.org/10.1109/DSN.2009.5270346 (letzter Aufruf: 28.06.2021)

Softmax-Werten einer KI-Anwendung zur Gesichtserkennung ein Trainingsbild einer Person zu rekonstruieren[83]. Methoden, um von einem Modell auf Rohdaten zurückzuschließen, die in aller Regel nicht preisgegeben werden sollen, werden als Modellinversion bezeichnet und sind Gegenstand aktueller Forschung.

Eine Möglichkeit, große Datenmengen und Modelle zu schützen, ist das *Federated Learning*. Da beim *Federated Learning* ausschließlich Modellparameter ausgetauscht werden, muss müssen potenzielle Angreifer*innen sehr viel Energie aufwenden, um auf die Trainingsdaten zurückschließen zu können. Zunächst müsste der*die Angreifer*in an die Modellparameter gelangen. Sie liegen bei dem*der Koordinator*in und den lokalen Agent*innen und werden zwischen ihnen ausgetauscht. Die Übertragung kann durch Verschlüsselung geschützt werden. Virtuelle Datenräume gewährleisten einen besonders sicheren Austausch von verteilten Daten und außerdem eine differenzierte Kontrolle über ihre Nutzung.

Im traditionellen Sinne ist die Vertraulichkeit von Daten ein Schutzziel der klassischen IT-Sicherheit. Während der Prüfkatalog die weiteren Schutzziele Integrität und Verfügbarkeit im traditionellen Sinne als Risikogebiete der **Dimension: Sicherheit (SI)** zuordnet, bemisst er dem Datenschutz hingegen eine eigene Dimension zu. Dies rührt aus der Tatsache, dass sich aufgrund der Eigenschaften von ML-Verfahren neuartige Risiken bezüglich der Vertraulichkeit ergeben, die über das reine Abgreifen gespeicherter Daten – wie es in der klassischen IT-Sicherheit behandelt wird – hinausgehen. Wie in den vorigen Absätzen beschrieben, können KI-Technologien Daten intelligent verknüpfen und derart etwa Personenbezüge herstellen bzw. rekonstruieren. Hinzu kommt das Risiko der Modellextraktion, sowie das Risiko der Extraktion von Trainingsdaten aus dem Modell. Insgesamt eröffnet der Datenschutz im Kontext von KI-Anwendungen eine eigene Risikolandschaft, die neue, KI-spezifische Maßnahmen erfordert und dementsprechend als eigenständige Dimension zu betrachten ist.

Nichtsdestotrotz steht der Datenschutz weiterhin im direkten Zusammenhang mit Risiken bezüglich der Integrität einer KI-Anwendung. Dementsprechend sollten zum Schutz von Daten ggf. auch die im **Risikogebiet: Integrität und Verfügbarkeit (IV)** aufgezeigten Maßnahmen der klassischen IT-Sicherheit wie etwa Verschlüsselung oder Identitäts- und Berechtigungsmanagement in Betracht gezogen werden. Die Maßnahmen in der Dimension Datenschutz hingegen fokussieren auf etablierte KI-spezifische Methoden. Aufgrund der Fülle möglicher Verfahren kann diesbezüglich jedoch nur eine Auswahl dargestellt werden. Insbesondere sind auch neue, nicht im Prüfkatalog aufgeführte Maßnahmen zur Abschwächung von Risiken zulässig.

Ziel der Dimension Datenschutz ist es, sicherzustellen, dass Datenschutzrisiken und -Maßnahmen der KI-Anwendung, unter Berücksichtigung der besonderen Herausforderungen durch die Künstliche Intelligenz, so ausreichend analysiert und dokumentiert sind, dass die Datenschutzbeauftragten sinnvoll darin unterstützt werden, die Untersuchung und letztlich die Entscheidung bzgl. der Datenschutz-Freigabe durchzuführen.

Die Risikogebiete in der Dimension Datenschutz sind die folgenden:

1. **Schutz personenbezogener Daten:** Dieses Risikogebiet behandelt Risiken, die mit der nicht DSGVO-konformen Nutzung personenbezogener Daten durch die KI-Anwendung verbunden sind, sowie das Risiko der Re-Identifikation von Personen in einem Datensatz.

2. **Schutz geschäftsrelevanter Information:** Dieses Risikogebiet adressiert Risiken, die dadurch entstehen, dass geschäftsrelevante Informationen durch die KI-Anwendung unerwünscht preisgegeben werden.

3. **Beherrschung der Dynamik:** Dieses Risikogebiet behandelt die Risiken, dass neue Hintergrundinformationen etwa zur Erstellung eines Personenbezugs entstehen, oder, dass sich die Anforderungen an die Verarbeitung von Daten durch die KI-Anwendung ändern.

---

83 Fredrikson, M., Jha, S., & Ristenpart, T. (2015). Model inversion attacks that exploit confidence information and basic countermeasures. Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, 1322–1333. https://doi.org/10.1145/2810103.2813677 (letzter Aufruf: 28.06.2021)

Zu der Aufteilung in Risikogebiete ist abschließend anzumerken, dass sich das **Risikogebiet: Schutz personenbezogener Daten (PD)** und das **Risikogebiet: Schutz geschäftsrelevanter Information (GI)** nur geringfügig in den dort aufgeführten, potenziell ergreifbaren Maßnahmen unterscheiden. Die Unterteilung dieser Themen in zwei Risikogebiete wird dennoch vorgenommen, da sich die Schutzanforderungen in Bezug auf personen- und geschäftsbezogene Daten stark unterscheiden können. Sollten Wiederholungen aufkommen, können diese durch Verweise auf bestehende Dokumentationen aus dem jeweils anderen Risikogebiet vermieden werden.

Schutzbedarfsanalyse

Obwohl sich bei personen- und geschäftsbezogenen Daten in der Regel verschiedene Schadensszenarien ergeben, hängt die Höhe potenzieller Schäden dennoch in beiden Fällen von der Art oder Kategorie der durch die KI-Anwendung verarbeiteten bzw. gespeicherten Daten ab.

Der Umgang mit personenbezogenen Daten ist durch die europäische Datenschutzgrundverordnung sowie das Bundesdatenschutzgesetz geregelt. Hierbei ist zu beachten, dass nicht erst durch einen unberechtigten Zugriff Dritter gegen die gesetzlichen Vorgaben verstoßen wird, sondern diese beispielsweise auch bei bloßem Vorliegen unzulässiger Zugriffsmöglichkeiten, unangemessen langer Speicherung oder Auskunftsunfähigkeit über die gespeicherten Daten verletzt werden. Sowohl der immaterielle Schaden durch Verletzung der Persönlichkeitsrechte der betroffenen Person(en) sowie die Höhe eines potenziellen finanziellen Schadens, z. B. durch Bußgelder oder Rufschädigung, ist abhängig von der Bedeutung/Kategorie der gespeicherten personenbezogenen Daten.

Analog ist der wirtschaftliche Schaden einer Organisation bzw. eines Unternehmens im Fall unberechtigten Zugriffs auf lizenzgebundene Daten oder Bekanntwerden von Geschäftsgeheimnissen abhängig von deren Art und Bedeutung. Der Begriff »lizenzgebundene Daten« wird im Folgenden vereinfachend für alle Arten von Daten verwendet, an denen Rechte Dritter vorliegen. Allgemeiner werden als geschäftsbezogene Daten solche Daten bezeichnet, die Informationen etwa über den*die Betreiber*in, insbesondere Geschäftsgeheimnisse, enthalten.

Der Schutzbedarf wird folgendermaßen kategorisiert:

Hoch	Der Schutzbedarf wird als hoch eingestuft, wenn eines der folgenden drei Szenarien zutrifft: Es werden personenbezogene Daten verarbeitet, die besonders sensible persönliche Informationen enthalten, oder deren Bekanntwerden für die entsprechende Person wirtschaftliche oder sicherheitskritische Konsequenzen hätte. **Beispiel:** Patientenakte, Führungszeugnis, Kontodaten, Bewerbungsunterlagen Es werden lizenzgebundene Daten verarbeitet, bei deren Bekanntwerden/Zugriff durch Dritte vertragliche Vereinbarungen verletzt würden. **Beispiel:** Daten von anderen Unternehmen wurden eingekauft, um das Modell zu trainieren. Zugriff auf diese Daten durch Dritte würde gegen die Vertragsvereinbarungen verstoßen. Es werden organisations-/geschäftsbezogene Daten verarbeitet, bei deren Bekanntwerden/Zugriff durch Dritte die Integrität oder Wettbewerbsfähigkeit der Organisation stark geschädigt würde. **Beispiel:** Modellextraktion hätte zur Folge, dass die entsprechende KI-Anwendung durch andere Organisationen kopiert oder gezielt manipuliert werden könnte.

Mittel	Der Schutzbedarf wird als mittel bewertet, wenn eines der folgenden drei Szenarien zutrifft und für keine der genannten Datenkategorien (personen-/geschäftsbezogen oder lizenzgebunden) ein gemäß der oberen Zeile dieser Tabelle hohes Schadenspotenzial besteht. Die KI-Anwendung verarbeitet/speichert ausschließlich solche Daten, die weder sensible persönliche Informationen enthalten, noch bei Zugriff durch Dritte einen großen wirtschaftlichen Nachteil oder die Gefährdung der Sicherheit der betroffenen Person zur Folge hätten. Die KI-Anwendung verarbeitet/speichert lizenzgebundene Daten, bei deren Zugriff Dritter vernachlässigbare Konsequenzen folgen könnten. Die KI-Anwendung verarbeitet/speichert geschäftsbezogene Daten, deren Bekanntwerden einen mittleren, nicht existenzbedrohenden wirtschaftlichen Schaden zur Folge haben könnte. **Beispiel:** Freizeitinteressen einer Person, abgespielte Songs, aufgerufene Videos in anonymisierter Form **Beispiel:** Eine KI-Anwendung, die auf Grundlage von öffentlich zugänglichen *Social Media*-Daten eine Trendanalyse vornimmt.
Gering	Durch die KI-Anwendung werden personenbezogene Daten weder abgefragt noch verarbeitet oder gespeichert. Außerdem speichert/verarbeitet die KI-Anwendung keine lizenzgebundenen Daten. Das Bekanntwerden der verarbeiteten Daten sowie Eigenschaften des Modells (z. B. Modellparameter) hätte keine oder lediglich vernachlässigbare Auswirkungen auf die Integrität oder Wettbewerbsfähigkeit der Organisation. **Beispiel:** Ein Unternehmen setzt eine Standard-KI-Lösung für die Prognose der Marktentwicklung ein. In der Branche haben andere Unternehmen vergleichbare Lösungen, und es wird vermutet, dass seitens der Konkurrenz kein Anreiz besteht, dieses System auszuspähen oder zu kopieren. Es werden z. B. Daten aus dem DAX oder andere Wirtschaftskennzahlen eingesetzt, die frei verfügbar sind.

[DS-S] Dokumentation der Schutzbedarfsanalyse

Anforderung: Do

- Der Schutzbedarf der KI-Anwendung in der Dimension Datenschutz wird als *gering, mittel* oder *hoch* festgelegt. Die Wahl der Kategorie *gering/mittel/hoch* wird unter Bezugnahme auf die oben angeführte Tabelle ausführlich begründet.

Falls der Schutzbedarf für die Dimension Datenschutz *gering* ist, so ist keine nähere Betrachtung der einzelnen Risikogebiete erforderlich. Wurde hingegen ein *mittlerer* oder *hoher* Schutzbedarf ermittelt, so muss im Folgenden jedes Risikogebiet genauer untersucht werden.

9.1 Risikogebiet: Schutz personenbezogener Daten (PD)

Verarbeitet eine KI-Anwendung personenbezogene Daten, so besteht das Risiko, dass durch spezielle KI-Verfahren oder unter Hinzunahme von Hintergrundwissen (z. B. anderen Datensätzen) Personen aus dem Datenbestand re-identifizierbar sind. Je nach Sensibilität der Informationen, die derart über eine Person bekannt werden, bedeutet dies eine schwere Verletzung ihrer Persönlichkeitsrechte. Hieraus ergibt sich die Anforderung, dass die durch die KI-Anwendung abgefragten, verarbeiteten oder gespeicherten Daten sowohl während des Trainings als auch im Betrieb wirksam geschützt werden müssen.

Gemäß der Datenschutzgrundverordnung (DSGVO) dürfen KI-Anwendung auf personenbezogene Daten ausschließlich mit Einwilligung der Betroffenen Zugriff nehmen. Eine Weiterverarbeitung sowie die Weitergabe an Dritte dürfen – vorbehaltlich weiterer Beschränkungen – ausschließlich mit Zustimmung der Rechtsgutsinhaber*innen erfolgen. Es muss sichergestellt werden, dass keine Schutzlücken bestehen, die einen unberechtigten Zugriff ermöglichen. Der einzelnen Person muss die Möglichkeit der Löschung ihrer Daten eingeräumt werden.[84]

In Bezug auf den Schutz personenbezogener Daten garantiert die DSGVO betroffenen Personen außerdem ein weitreichendes und jederzeitiges Widerspruchsrecht gegen die Verarbeitung ihrer Daten. Die rechtskonforme Umsetzung speziell eines nachträglichen Widerspruchs kann im Zusammenhang mit KI-Anwendungen besondere technische und organisatorische Herausforderungen mit sich bringen. Welche Pflichten sich konkret für Betreiber*innen aus der DSGVO im Zusammenhang mit der Löschung personenbezogener Daten ergeben, ist zwar in rechtlicher Hinsicht nicht vollständig geklärt. Eine mögliche Nachweispflicht darüber, dass eine Löschung personenbezogener Daten vollständig erfolgt ist, sowie die damit verbundenen, potenziell aufwändigen technischen und organisatorischen Konsequenzen können jedoch durch geeignete Maßnahmen vermieden werden. Beispielsweise kann die Anonymisierung von Trainingsdaten weitestgehend verhindern, dass Personenbezüge erstellt werden. Dadurch würde ein ggf. erforderlicher Nachweis, dass nach Löschung der Daten kein von einem Widerspruch betroffener Personenbezug mehr besteht, und insbesondere der Extremfall, das Modell im Fall eines Widerspruchs komplett neu anlernen zu müssen, umgangen.

Zu den erforderlichen Maßnahmen gehört ferner, dass Betroffene über den Zweck und Einsatz ihrer personenbezogenen oder daraus abgeleiteten Daten informiert werden. Zusätzlich zu den bereitzustellenden Einwilligungs-, Auskunfts-, Einspruchs- und Widerrufsmechanismen bzgl. der Nutzung personenbezogener Daten sind die Grundsätze der Datensparsamkeit und zweckgebundenen Verwendung einzuhalten.

Mit der im Folgenden beschriebenen Risikoanalyse soll ermittelt werden, welche spezifischen Gefährdungen des Schutzes personenbezogener Daten für die zu untersuchende KI-Anwendung möglich sind. Dazu soll insbesondere untersucht werden, von welcher Art und Bedeutung die im Zusammenhang mit der KI-Anwendung abgefragten bzw. gespeicherten Daten sind und wo potenzielle Schutzlücken bestehen. Insbesondere sollte unter Berücksichtigung der ggf. eingesetzten Maßnahmen zur Anonymisierung oder Aggregation von Daten ein geringes Risiko der Re-Identifikation bzw. der Herstellbarkeit eines Personenbezugs durch Verknüpfung mit Hintergrundwissen erzielt werden. Zwar ist es in der Regel nicht möglich, die Wiederherstellung eines Personenbezugs vollständig auszuschließen, jedoch sollte der Aufwand der Re-Identifizierung von Personen in einem Datensatz unverhältnismäßig hoch sein.

84 Die Darstellung in diesem Abschnitt sowie teils in den folgenden Abschnitten ist stark angelehnt an das Kapitel »3.6 Datenschutz« des Whitepapers: Poretschkin, M., et al. (2019). Vertrauenswürdiger Einsatz von Künstlicher Intelligenz. Sankt Augustin: Fraunhofer-Institut für Intelligente Analyse und Informationssysteme IAIS. https://www.ki.nrw/wp-content/uploads/2020/03/Whitepaper_KI-Zertifizierung.pdf (letzter Aufruf: 18.06.2021)

9.1.1 Risikoanalyse und Zielvorgaben

[DS-R-PD-RI-01] Risikoanalyse der Trainingsdaten

Anforderung: Do

- Es werden die Eigenschaften der verwendeten Trainingsdaten beschrieben und deren Wahl bzw. Eignung dokumentiert und begründet. Weiterhin wird erläutert, ob die Trainingsdaten der KI-Komponente Informationen beinhalten, die einen Personenbezug ermöglichen. Dabei sind insbesondere die Attribute, die Menge, sowie die Verknüpfungsmöglichkeiten der Daten mit anderen (personenbezogenen) Hintergrundinformationen zu dokumentieren.
- Es liegen Beispiel-Trainingsdaten vor, anhand derer die in der Dokumentation beschriebenen Eigenschaften der Trainingsdaten hinsichtlich des Schutzes personenbezogener Daten nachvollzogen werden können.

[DS-R-PD-RI-02] Risikoanalyse der Input- und Nutzungsdaten

Anforderung: Do

- Es wird dokumentiert und erläutert, welche der Input- bzw. Nutzungsdaten, die während des Betriebs der KI-Anwendung erfasst und gespeichert werden, einen potenziellen Personenbezug ermöglichen. Dabei wird insbesondere beschrieben, welche potenziell personenbezogenen Attribute vorliegen und welche Möglichkeiten zur Verknüpfung mit anderen Datensätzen bestehen.
- Außerdem wird dargelegt, in welcher Menge potenziell personenbezogene Ein- und Ausgaben sowie weitere Nutzungsdaten (z. B. durch *Logging*) erhoben und gespeichert werden.
- Zuletzt ist anzugeben, welche der erhobenen Daten als Trainingsdaten genutzt werden sollen (siehe **[DS-R-PD-RI-01]**), oder nur zu Auskunfts- und Nachweiszwecken oder zu weiteren Zwecken (z. B. zur Lastanalyse) gespeichert und genutzt werden.
- Es liegen Beispieldaten vor, anhand derer die in der Dokumentation beschriebenen Eigenschaften der Eingabe- und Nutzungsdaten hinsichtlich des Schutzes personenbezogener Daten nachvollzogen werden können.

[DS-R-PD-RI-03] Biometrische Merkmale

Anforderung: Do

- Es wird dokumentiert, inwiefern biometrische Daten (z. B. Bilder, Handschrift, Gesundheitsdaten, Fingerabdrücke, Tasten- und Mausbedienung, …) durch die KI-Anwendung erhoben und genutzt werden. Insbesondere wird erläutert, ob und mit welchem Hintergrund-Wissen und KI-Verfahren aus diesen Daten ein Personenbezug erstellt werden könnte.

[DS-R-PD-RI-04] Modellergebnisse und Seitenkanäle

Anforderung: Do

- Es wird analysiert und dokumentiert, inwiefern die Ergebnisse der KI-Anwendung für die Erstellung eines ungewollten Personenbezugs anfällig sind. Dazu zählt neben der Analyse der reinen Ausgabe des Modells auch die Analyse von Verknüpfungsmöglichkeiten der Ausgabe mit Hintergrundinformation. Ferner sollte die KI-Anwendung im Hinblick auf mögliche Seitenkanäle untersucht werden. Beispielsweise könnte die Verarbeitungszeit von Input-Daten einen Rückschluss auf personenbezogene Informationen ermöglichen.

[DS-R-PD-RI-05] Risikoanalyse der KI-Anwendung insgesamt

Anforderung: Do

- **Risikoanalyse:** Es wird dokumentiert, welche Gesamtrisiken und potenzielle Schäden im Zusammenhang mit den in **[DS-R-PD-RI-01]** bis **[DS-R-PD-RI-04]** beschriebenen, potenziell personenbezogenen Daten sowohl bei der Entwicklung als auch beim Betrieb der KI-Anwendung bestehen. Dabei wird zum einen analysiert, welche Risiken im Hinblick auf unbeabsichtigte bzw. unerlaubte Verarbeitungs- und Zugriffsmöglichkeiten auf personenbezogene Daten bestehen, die konform zu den Datenschutzvorgaben erhoben und verwendet wurden. Zum anderen wird das Risiko untersucht, dass unter Nutzung von Hintergrundinformationen und/

oder unter Nutzung von KI-Techniken unerwünscht bzw. unerlaubt Personenbezüge erstellt werden können. Ferner wird abgeschätzt, welche Schäden resultieren können, falls die potenziell personenbezogenen Daten nicht bestimmungsgemäß eingesehen und genutzt werden.

- **Zielvorgaben:** Ausgehend von der Risikoanalyse werden formale Schutzziele im Zusammenhang mit personenbezogenen Daten formuliert, die mindestens konform mit der DSGVO sein müssen und im Folgenden genauer ausgeführt und geprüft werden.

9.1.2 Kriterien zur Zielerreichung

[DS-R-PD-KR-01] Quantifizierung des Datenschutz-Risikos

Anforderung: Do

- Es wird dokumentiert, gemäß welchen Kriterien das Risiko der Erstellung eines Personenbezugs beurteilt werden soll. Kriterien sollten vorzugsweise quantitativer Natur sein, wie beispielsweise
  - Gruppengrößen beim Einsatz von Anonymisierungsverfahren (z. B. K-*Anonymity*, L-*Diversity* oder T-*Closeness*),
  - Gruppengrößen bei statistischer Aggregation,
  - Privatsphäre-Budgets (*Differential Privacy*) oder Beschränkungen von Abfragemöglichkeiten,
  - Aufwand für das Heranziehen nützlicher Hintergrundinformationen,
  - Aufwand für das Erstellen eines Personenbezugs über Berechnungen (z. B. beim Einsatz von Verschlüsselungsverfahren).

  Zusätzlich können qualitative Kriterien in Betracht gezogen werden, wie etwa der Nachweis von berechtigtem Interesse im Sinne der DSGVO sowie DSGVO-konformer Umgang mit der expliziten Einwilligung zur Nutzung personenbezogener Daten. Darüber hinaus können auch andere, hier nicht aufgeführte Kriterien festgelegt werden. Falls für verschiedene Datensätze unterschiedliche Kriterien festgelegt werden, so ist dies pro Datensatz ausführlich zu begründen.
- Für die gewählten quantitativen Kriterien werden zudem Zielintervalle festgelegt, bei deren Erreichen ein vertretbares Risiko hergestellt ist.
- Die Wahl der Kriterien und ggf. Zielintervalle wird nachvollziehbar begründet. Dabei wird insbesondere verdeutlicht, inwiefern diese mit den in **[DS-R-PD-RI-05]** festgelegten Zielvorgabenkonform sind.

9.1.3 Maßnahmen

9.1.3.1 Daten

[DS-R-PD-MA-01] Anonymisierung

Anforderung: Do

- Es wird dokumentiert, welche Verfahren zur Anonymisierung von Daten eingesetzt werden. Unter der Vielzahl von Verfahren zur Anonymisierung sind beispielsweise etabliert:
  - K-*Anonymity*
  - *Differential Privacy*

Die Wahl der eingesetzten Verfahren zur Anonymisierung ist zu begründen. Außerdem ist die Wirksamkeit der Verfahren zu bewerten, u. a. im Hinblick auf potenziell verfügbare Hintergrundinformationen.

[DS-R-PD-MA-02] Pseudonymisierung

Anforderung: Do

- Es wird dokumentiert, welche Verfahren zur Pseudonymisierung personenbezogener Daten (wie etwa *hashing*) zum Einsatz kommen, um die Re-Identifizierung von Personen in einem Datensatz zu erschweren.
- Pseudonymisierung ist nicht der Anonymisierung gleichzusetzen und bietet in der Regel keinen ausreichenden Schutz im Hinblick auf die Erstellung eines Personenbezugs. Daher ist darzulegen, inwiefern die eingesetzten Verfahren zur Pseudonymisierung in Kombination mit weiteren ergriffenen Maßnahmen (z. B. aus dem **Risikogebiet: Integrität und Verfügbarkeit (IV)** der Dimension Sicherheit) wirksam sind, und ggf., welche Lücken bestehen.

[DS-R-PD-MA-03] Perturbation von Daten zur Modellbildung

Anforderung: Do

- Es wird dokumentiert, in welcher Weise Daten ggf. durch Hinzufügen von absichtlichen Zufallsverzerrungen zur Modellbildung verändert (perturbiert) werden, um die Extraktion personenbezogener Daten zu verhindern oder zu erschweren. Mögliche Verfahren der Perturbation sind beispielsweise:
  – Hinzufügen von additivem oder multiplikativem Zufallsrauschen. Die Art (z. B. weißes oder uniformes Rauschen) und Stärke (z. B. Amplitude und Standardabweichung) des Rauschens sind anzugeben.
  – Zufälliges Durchmischen (*Shuffling*) von Attributwerten.

  Sollten andere Methoden zum Einsatz kommen, so ist zu begründen, warum sie geeignet sind, die Extraktion personenbezogener Daten zu erschweren.

[DS-R-PD-MA-04] Aggregation und Generalisierung von Daten zur Modellbildung

Anforderung: Do

- Es wird dokumentiert, inwiefern Daten bei der Modellbildung verknüpft bzw. aggregiert und generalisiert werden, um die Extraktion personenbezogener Daten zu verhindern oder zu erschweren. Ferner wird die Wirksamkeit der ergriffenen Aggregationen bzw. Generalisierungen hinsichtlich des Datenschutzes bewertet.

9.1.3.2 KI-Komponente

[DS-R-PD-MA-05] Datensparsamkeit zur Modellbildung

Anforderung: Do

- Es wird dokumentiert und begründet, dass die durchgeführte Modellbildung nicht auch unter Nutzung anderer Daten mit geringerer Sensibilität (hinsichtlich eines Personenbezugs) möglich ist.

[DS-R-PD-MA-06] Zweckgebundenheit der KI-Anwendung

Anforderung: Do

- Falls die KI-Anwendung personenbezogene Daten auf Basis einer zweckgebundenen Einwilligung verarbeitet, wird dokumentiert, dass das implementierte ML-Modell diese Daten nur gemäß dem bewilligten Zweck nutzt. Hierzu kann ggf. auf Dokumentationen aus der **Dimension: Verlässlichkeit (VE)** verwiesen werden.

[DS-R-PD-MA-07] Neuheit der Ausgaben

Anforderung: Do

- Insbesondere für generative Modelle, aber auch für prädiktive Modelle etwa zur Vervollständigung von Eingaben, besteht prinzipiell das Risiko, dass die Ausgaben des ML-Modells Teile der Trainingsdaten wiedergeben. Falls die Trainingsdaten gemäß **[DS-R-PD-RI-01]** potenziell einen Personenbezug ermöglichen, sollte je nach Modelltyp sichergestellt werden, dass die Ausgaben der KI-Komponente hinreichend stark von den Trainingsdaten abweichen und diese nicht in unverhältnismäßigem Ausmaß unbeabsichtigt preisgeben. In

diesem Fall ist zu dokumentieren, welche Maßnahmen ergriffen wurden, um das unmittelbare Preisgeben potenziell personenbezogener Trainingsdaten durch Ausgaben der KI-Anwendung zu verhindern. Dabei kann unter anderem auf ergriffene Maßnahmen zur Datenvorverarbeitung wie etwa **[DS-R-PD-MA-01]** bis **[DS-R-PD-MA-04]** sowie auf Dokumentationen aus dem **Risikogebiet: Integrität und Verfügbarkeit (IV)** der Dimension Sicherheit verwiesen werden, beispielsweise die Beschränkung der Abfragemöglichkeiten in **[SI-R-IV-MA-10]** gemäß **[SI-R-IV-KR-03]**.

- Ferner wird das Risiko des unmittelbaren Preisgebens potenziell personenbezogener Trainingsdaten durch die KI-Anwendung, u. a. angesichts der Dimensionalität der Einbettung etwa bei generativen Modellen, bewertet.

[DS-R-PD-MA-08] *Federated Learning*

Anforderung: Do

- Eine Möglichkeit, den unerwünschten Zugriff auf Daten zu erschweren, ist das verteilte Lernen bzw. *Federated Learning*. Hierbei werden Modelle lokal an verschiedenen Rechnerknoten trainiert, sodass die jeweiligen Trainingsdaten ihre lokale Position nicht verlassen müssen. Die separat erzeugten Modelle werden anschließend zu einem globalen Modell zusammengefügt. Wurde das ML-Modell durch *Federated Learning* (oder eine Abwandlung davon) gebildet, so ist dies zu dokumentieren. Insbesondere ist zu beschreiben, inwiefern durch das verteilte Lernen der KI-Komponente das Ausspähen personenbezogener Daten erschwert wird.

9.1.3.3 Einbettung

[DS-R-PD-MA-09] Ungewollter Abfluss von Informationen

Anforderung: Do

- Durch gezieltes Abfragen der KI-Anwendung können unter Umständen Daten, insbesondere Trainingsdaten des ML-Modells, unmittelbar oder über Seitenkanäle rekonstruierbar sein. Beispielsweise könnte durch abwechselndes Abfragen der KI-Anwendung und Anpassen des Inputs ein Input-Datum konstruiert werden, das eine bestimmte Ausgabe erzielt. Die Gestalt des Input-Datums könnte wiederum einen Rückschluss auf die Trainingsdaten oder auf durch das ML-Modell gelernte Zusammenhänge geben. Falls gemäß **[DS-R-PD-RI-04]** das Risiko besteht, dass auf diese Weise personenbezogene Daten ausgespäht werden, ist zu dokumentieren, welche Maßnahmen ergriffen wurden, um den ungewollten Abfluss von Informationen durch Abfragen der KI-Anwendung oder über Seitenkanäle zu verhindern bzw. zu erschweren. Hierbei kann unter anderem auf ergriffene Maßnahmen zur Datenvorverarbeitung wie etwa **[DS-R-PD-MA-01]** bis **[DS-R-PD-MA-04]** sowie auf Dokumentationen aus dem **Risikogebiet: Integrität und Verfügbarkeit (IV)** der Dimension Sicherheit verwiesen werden, beispielsweise die Beschränkung der Abfragemöglichkeiten in **[SI-R-IV-MA-10]** gemäß **[SI-R-IV-KR-03]**.

9.1.3.4 Maßnahmen für den Betrieb

[DS-R-PD-MA-10] Speicherung und Löschung

Anforderungen: Do | Pr

- Die technische Umsetzung und der Speicherort potenziell personenbezogener Daten (Trainings-, Input-, Ausgabe- und Nutzungsdaten) werden ebenso dokumentiert wie die ergriffenen Maßnahmen, um gespeicherte personenbezogene Daten vor informationstechnischen Angriffen zu schützen. Hierbei kann ggf. auf Maßnahmen aus dem **Risikogebiet: Integrität und Verfügbarkeit (IV)** der Dimension Sicherheit verwiesen werden, wie beispielsweise Maßnahmen zur Vertraulichkeit (siehe **[SI-R-IV-MA-03]**) sowie zur Sicherung von Daten (siehe **[SI-R-IV-MA-04]**) und des Speicherorts (siehe **[SI-R-IV-MA-05]**).
- Außerdem werden die technischen Verfahren zur Löschung potenziell personenbezogener Daten beschrieben. Insbesondere wird dargelegt, wie damit umgegangen wird, dass Personen unter Umständen die Einwilligung zur Verarbeitung ihrer Daten widerrufen können. Falls in diesem Kontext betriebliche Prozesse etabliert wurden, sind diese zu dokumentieren.

[DS-R-PD-MA-11] Auskunftsfähigkeit bzgl. personenbezogener Daten

Anforderungen: Do | Pr

- Es wird dokumentiert, wie sichergestellt ist, dass Betroffene und Nutzer*innen Auskunft über die von ihnen verwendeten bzw. über sie erhobenen Daten erhalten können.
- Es wird dokumentiert, wie Betroffene und Nutzer*innen erfahren können, welche Entscheidungen die KI-Anwendung bzgl. ihrer Person oder ihrer Anfragen getroffen hat. Dabei kann u. U. auf Dokumentationen aus der **Dimension: Transparenz (TR)** und/oder **Dimension: Autonomie und Kontrolle (AK)** verwiesen werden.
- Eine Verletzung der Vertraulichkeit von Daten kann, je nach Art, ggf. nicht wiederhergestellt werden. Zur Schadensreduktion können jedoch Vorgehensweisen, beispielsweise zur Benachrichtigung von Betroffenen, eingeführt werden. Wurden solche Prozesse etabliert, sind diese zu dokumentieren.

9.1.4 Gesamtbewertung

[DS-R-PD-BW-01] Bewertung der Anonymisierung

Anforderung: Do

- Es wird erläutert, inwiefern die ergriffenen und dokumentierten Maßnahmen zur Anonymisierung darin resultieren, dass die anonymisierten Daten keine unerwünschte/unautorisierte Erstellung eines Personenbezugs erlauben bzw. dies nur unter sehr hohem Aufwand, der typischerweise in keiner Relation zum erwartbaren Nutzen steht, möglich ist.

[DS-R-PD-BW-02] Erklärung zur Datenschutz-Konformität

Anforderung: Do

- Es wird dokumentiert, dass die Nutzung der in **[DS-R-PD-RI-01]** bis **[DS-R-PD-RI-04]** beschriebenen, potenziell personenbezogenen Daten mit der DSGVO und dem BDSG verträglich ist. Dabei ist unter anderem auf die Risikoabschätzung, die ergriffenen Einwilligungs- und Schutzmechanismen sowie das berechtigte Interesse Bezug zu nehmen.
- Zudem liegt eine Datenschutzfolgeabschätzung im Sinne des Art. 35 Abs. 1 DSGVO vor, oder aber es wird aussagekräftig begründet, dass diese aufgrund der ergriffenen Maßnahmen (z. B. zur Anonymisierung von Daten) nicht erforderlich ist. Wurde eine Datenschutzfolgeabschätzung erstellt, so ist deren Konformität mit der DSGVO durch den*die Datenschutzbeauftragte*n bestätigt.
- Abschließend wird zusammenfassend begründet, dass die in **[DS-R-PD-KR-01]** gesetzten Ziele erreicht wurden.
- Sofern nicht alle in **[DS-R-PD-KR-01]** spezifizierten Anforderungen erfüllt werden, sind die Abweichungen zu dokumentieren. Dies gilt ebenfalls für nur teilerfüllte Anforderungen, bei denen etwa die Kriterien nicht, oder nicht immer, erreicht wurden.

9.2 Risikogebiet: Schutz geschäftsrelevanter Information (GI)

Abgesehen von personenbezogenen Daten entstehen im Zuge der Digitalisierung von Wirtschaftsprozessen immer mehr schützenswerte Geschäftsdaten. Auf leistungsstarken Rechnern werden mithilfe großer Datenmengen Modelle des Maschinellen Lernens für zahlreiche Anwendungen trainiert: von der Anomalieerkennung im *Condition Monitoring* und der präventiven Wartung (*Predictive Maintenance*) über Empfehlungen zur Maschineneinstellung bis hin zu autonomen Fahrzeugen, kooperativen Robotern und intelligenten Steuerungen. Das Lernen auf diesen großen Datenmengen findet zurzeit vorwiegend in der Cloud statt, also auf einer zentralen *Big Data*-Plattform, auf der historische Daten kontinuierlich durch Datensätze ergänzt werden, die im Einsatz neu gewonnen werden. Diese Art der Umsetzung ist in zahlreichen Bereichen jedoch weder technisch wünschenswert noch rechtlich möglich: Während im Gesundheitswesen der Datenschutz die primäre Hürde darstellt, bergen die in Geschäftsdaten enthaltenen Informationen, etwa über Geräte und Maschinenproduzent*innen, das Risiko, dass Interna und Betriebsgeheimnisse ausgespäht werden.

Somit können auch nicht-personenbezogene Daten schützenswert sein, z. B. weil sie aus Wettbewerbsgründen geheim gehalten werden sollen, oder weil ihre Nutzung vertraglich geregelt ist. Letztere werden im Folgenden vereinfachend als lizenzgebundene Daten bezeichnet – d. h. alle Arten von Daten, an denen Rechte Dritter vorliegen. Dies können insbesondere auch öffentlich zugängliche Daten sein. Die Rechte der Dritten werden im Folgenden der Einfachheit halber als »Lizenzbedingungen« bezeichnet. Mit geschäftsrelevanter Information sind im Folgenden all jene Daten gemeint, die sich auf den*die Betreiber*in bzw. dessen Geschäfte und/oder Geschäftskontakte beziehen, und die vor dem Zugriff bzw. der unkontrollierten Einsehbarkeit durch Dritte geschützt werden sollen. Dazu zählen insbesondere Daten, die Geschäftsgeheimnisse enthalten, sowie ggf. lizenzgebundene Daten. Insbesondere fallen unter geschäftsrelevante Informationen all jene Daten, deren Bekanntwerden für das Unternehmen eine Minderung der Wettbewerbsfähigkeit zur Folge hätte oder der Sicherheit bzw. Integrität des*der Betreiber*in schaden würde. Auch das Modell in der KI-Komponente selbst kann eine geschäftsrelevante Information sein, wenn es beispielsweise ein Alleinstellungsmerkmal des*der Betreiber*in ist und diesem*dieser dadurch einen Wettbewerbsvorteil verschafft.

9.2.1 Risikoanalyse und Zielvorgaben

In der Risikoanalyse für dieses Risikogebiet soll ermittelt werden, inwiefern die zu prüfende KI-Anwendung geschäftsrelevante Information verarbeitet und welche potenziellen Gefährdungen des Risikogebiets im spezifischen Einsatzkontext möglich sind. Darauf basierend ist die Zielsetzung des Schutzes geschäftsrelevanter Information zu formulieren.

[DS-R-GI-RI-01] Risikoanalyse der Trainingsdaten

Anforderung: Do

- In Ergänzung zu und ggf. in Bezugnahme auf **[DS-R-PD-RI-01]** ist zu dokumentieren, ob und welche der genutzten Trainingsdaten geschäftsrelevante Informationen enthalten oder insbesondere lizenzgebunden sind.

[DS-R-GI-RI-02] Risikoanalyse der Modelleigenschaften

Anforderung: Do

- Es ist zu dokumentieren, ob und welche Eigenschaften des ML-Modells (z. B. die Art des Modells, Modellparameter) bzw. der KI-Komponente eine geschäftsrelevante Information darstellen oder insbesondere lizenzgebunden sind.

[DS-R-GI-RI-03] Risikoanalyse der Input- und Nutzungsdaten

Anforderung: Do

- In Ergänzung zu und ggf. in Bezugnahme auf **[DS-R-PD-RI-02]** ist zu dokumentieren, ob und welche der Input- und Nutzungsdaten geschäftsrelevante Informationen enthalten oder insbesondere lizenzgebunden sind.

[DS-R-GI-RI-04] Modellergebnisse und Seitenkanäle

Anforderung: Do

- In Ergänzung zu und ggf. in Bezugnahme auf **[DS-R-PD-RI-04]** ist die Sensibilität der Ergebnisse der KI-Anwendung bezüglich der Verletzung von Lizenzbedingungen bzw. der unerwünschten Einsehbarkeit geschäftsrelevanter Information zu analysieren. Dazu zählt neben der Analyse der reinen Ausgabe des Modells auch die Analyse von Verknüpfungsmöglichkeiten der Ausgabe mit Hintergrundinformation. Ferner ist die KI-Anwendung im Hinblick auf mögliche Seitenkanäle zu untersuchen.

[DS-R-GI-RI-05] Risikoanalyse der KI-Anwendung insgesamt

Anforderung: Do

- **Risikoanalyse:** Angesichts der in **[DS-R-GI-RI-01]** bis **[DS-R-GI-RI-04]** beschriebenen geschäftsbezogenen Daten wird analysiert, welche Risiken bezüglich des unberechtigten Zugriffs Dritter auf geschäftsrelevante Information und insbesondere bezüglich der Verletzung von Lizenzbedingungen sowohl bei der Entwicklung als auch beim Einsatz der KI-Anwendung bestehen. Dazu werden mögliche Gefährdungen der Vertraulichkeit spezifiziert, d. h. es werden für den vorliegenden Anwendungskontext plausible Ursachen bzw. Angriffsmöglichkeiten untersucht, durch die auf die beschriebenen sensiblen Daten unautorisiert zugegriffen werden könnte. Die Wahrscheinlichkeit, dass diese Gefährdungen eintreten, wird abgeschätzt. Ferner wird analysiert, welche potenziellen Schäden eine unautorisierte Einsichtnahme geschäftsrelevanter Information und insbesondere lizenzgebundener Daten zur Folge hätte.
- **Zielvorgaben:** Aufbauend auf der Risikoanalyse wird die Zielsetzung bezüglich des Schutzes geschäftsrelevanter Information und der Lizenzbedingungen formuliert. Diese muss mindestens die Einhaltung der mit den Daten verbundenen Lizenz- oder sonstigen Nutzungsbedingungen beinhalten.

9.2.2 Kriterien zur Zielerreichung

[DS-R-GI-KR-01] Quantifizierung des Risikos

Anforderung: Do

- Es werden Kriterien festgelegt und dokumentiert, anhand derer das Risiko, Lizenzbedingungen zu verletzen oder geschäftsrelevante Information nicht angemessen zu schützen, beurteilt werden soll. Falls für verschiedene Datensätze unterschiedliche Kriterien festgelegt werden, so ist dies pro Datensatz ausführlich zu begründen. Die Kriterien sollten vorzugsweise quantitativer Natur sein, es können aber auch qualitative Kriterien herangezogen werden. Bei der Wahl der Kriterien sollte die folgende Auflistung in Betracht gezogen werden:
  - Grad und Art der Veröffentlichung bzw. des unerwünschten Zugriffs auf Daten
  - Menge und Umfang der unerwünscht preisgegebenen Daten
  - Kritikalität der unerwünscht preisgegebenen Daten für die Geschäftsbeziehungen und Wettbewerbsfähigkeit des*der Betreiber*in
  - Kosten (z. B. bei Verletzung von Lizenzbedingungen)
  - Kritikalität der unerwünscht preisgegebenen Daten hinsichtlich der Manipulierbarkeit bzw. Angreifbarkeit der KI-Anwendung
  - Aufwand für das Heranziehen von Hintergrundinformationen

  - Falls für die Art der geschäftsbezogenen Daten relevant:
    - Gruppengrößen beim Einsatz von Anonymisierungsverfahren (z. B. K-*Anonymity*, L-*Diversity* oder T-*Closeness*)
    - Gruppengrößen bei statistischer Aggregation
    - Privatsphäre-Budgets (*Differential Privacy*) oder Beschränkungen bzgl. Abfragemöglichkeiten

  Die Wahl der Kriterien muss nachvollziehbar begründet werden. Insbesondere sind Kriterien, die nicht aus der obigen Auflistung stammen, zu erläutern.
- Für die gewählten quantitativen Kriterien werden zudem Zielintervalle festgelegt, bei deren Erreichen ein vertretbares Risiko hergestellt ist.
- Es wird dargelegt, dass die gewählten Kriterien und ggf. Zielintervalle die in **[DS-R-GI-RI-05]** definierten Zielvorgaben angemessen abbilden.

9.2.3 Maßnahmen

9.2.3.1 Daten

[DS-R-GI-MA-01] Perturbation von Daten zur Modellbildung

Anforderung: Do

- In Ergänzung zu **[DS-R-PD-MA-03]** wird dokumentiert, in welcher Weise Daten durch Hinzufügen von absichtlichen Zufallsverzerrungen bei der Modellbildung verändert (perturbiert) werden, um die Extraktion geschäftsrelevanter Information und insbesondere lizenzgebundener Daten zu verhindern oder zu erschweren. Mögliche Verfahren der Perturbation sind beispielsweise:
  - Hinzufügen von additivem oder multiplikativem Zufallsrauschen. Die Art (z. B. weißes oder uniformes Rauschen) und Stärke (z. B. Amplitude und Standardabweichung) des Rauschens sind anzugeben.
  - Zufälliges Durchmischen (*Shuffling*) von Attributwerten.

  Sollten andere Methoden zum Einsatz kommen, so ist zu begründen, warum sie geeignet sind, die Extraktion geschäftsrelevanter Information und insbesondere lizenzgebundener Daten zu erschweren.

[DS-R-GI-MA-02] Aggregationen und Generalisierung von Daten zur Modellbildung

Anforderung: Do

- In Ergänzung zu **[DS-R-PD-MA-04]** werden die zur Modellbildung vorgenommenen Verknüpfungen/Aggregationen und Generalisierungen von lizenzgebundenen Daten bzw. geschäftsrelevanter Information dokumentiert. Ferner werden die vorgenommenen Aggregationen und Generalisierungen dahingehend bewertet, inwiefern sie die ungewünschte Einsehbarkeit dieser Daten erschwert.

[DS-R-GI-MA-03] Anonymisierung

Anforderung: Do

- Falls dies für die Art der geschäftsrelevanten Informationen anwendbar und angemessen ist, so können in Ergänzung zu **[DS-R-PD-MA-01]** Mechanismen zur Anonymisierung auch für lizenzgebundene Daten bzw. geschäftsrelevante Information eingesetzt werden. Wurden solche Maßnahmen ergriffen, so ist zu erläutern, welches Verfahren zur Anonymisierung auf welchen geschäftsbezogenen Datensatz angewandt wurde. Außerdem ist die Wirksamkeit des Verfahrens u. a. im Hinblick auf potenziell verfügbare Hintergrundinformationen zu bewerten.

[DS-R-GI-MA-04] Pseudonymisierung

Anforderung: Do

- Falls dies für die Art der geschäftsrelevanten Informationen anwendbar und angemessen ist, so können in Ergänzung zu **[DS-R-PD-MA-02]** Mechanismen zur Pseudonymisierung auch für lizenzgebundene Daten bzw. geschäftsrelevante Information eingesetzt werden. Wurden solche Maßnahmen für geschäftsbezogene Daten ergriffen, so sind die Wahl des Verfahrens (wie etwa *Hashing*) sowie die damit verarbeiteten Datensätze zu dokumentieren.
- Pseudonymisierung ist nicht der Anonymisierung gleichzusetzen und bietet in der Regel keinen ausreichenden Schutz im Hinblick auf Re-Identifizierbarkeit, beispielsweise von pseudonymisierten Unternehmen. Daher ist darzulegen, inwiefern die eingesetzten Verfahren zur Pseudonymisierung in Kombination mit weiteren ergriffenen Maßnahmen (z. B. aus dem **Risikogebiet: Integrität und Verfügbarkeit (IV)**) wirksam sind, und ggf., welche Lücken bestehen.

[DS-R-GI-MA-05] *Data Obfuscation*

Anforderung: Do

- In Abschwächung und/oder Ergänzung zu **[DS-R-GI-MA-01]** kann der Inhalt von gespeicherten Daten verschleiert werden. Beispielsweise könnten Geburtsdaten nicht als Jahreszahlen sondern (unter Berücksichtigung des genauen Datums) als reelle Zahl im normalisierten Intervall (-1, +1) abgelegt werden. Dies kann dazu beitragen, dass bei Verlust der Daten ihr semantischer Kontext nicht erkannt wird und somit Deduktionen erschwert werden, ohne dass Zusammenhänge im Datensatz verfälscht werden. Es ist zu dokumentieren, ob, in welchem Umfang und in welcher Form eine Verschleierung von Daten vorgenommen wurde.

9.2.3.2 KI-Komponente

[DS-R-GI-MA-06] Zweckgebundenheit der KI-Anwendung

Anforderung: Do

- Im Fall lizenzgebundener Daten wird nachgewiesen und dokumentiert, dass die KI-Anwendung diese nur gemäß der bewilligten Lizenz nutzt. Hierzu kann u. U. auf Dokumentationen aus der **Dimension: Verlässlichkeit (VE)** verwiesen werden.
- Falls die KI-Anwendung geschäftsbezogene Daten verarbeitet, wird nachgewiesen und dokumentiert, dass die Verarbeitung der geschäftsrelevanten Informationen durch die KI-Anwendung tatsächlich notwendig für deren Funktionalität und Zweckmäßigkeit ist.

[DS-R-GI-MA-07] Neuheit der Ausgaben

Anforderung: Do

- Falls die Trainingsdaten gemäß **[DS-R-GI-RI-01]** schützenswerte geschäftsrelevante Informationen oder insbesondere lizenzgebundene Daten enthalten, ist analog zu **[DS-R-PD-MA-07]** zu dokumentieren, welche Maßnahmen ergriffen wurden, um zu verhindern, dass die Trainingsdaten durch Ausgaben der KI-Anwendung unmittelbar preisgegeben werden. Dabei kann ggf. direkt auf **[DS-R-PD-MA-07]** verwiesen werden, oder auf ergriffene Maßnahmen zur Datenvorverarbeitung wie etwa **[DS-R-GI-MA-01]** bis **[DS-R-GI-MA-04]** sowie auf Dokumentationen aus dem **Risikogebiet: Integrität und Verfügbarkeit (IV)** der Dimension Sicherheit, beispielsweise die Beschränkung der Abfragemöglichkeiten (siehe **[SI-R-IV-MA-10]**).
- Ferner wird das Risiko des unmittelbaren Preisgebens geschäftsbezogener Trainingsdaten durch die KI-Anwendung, u. a. angesichts der Dimensionalität der Einbettung etwa bei generativen Modellen, bewertet.

[DS-R-GI-MA-08] *Federated Learning*

Anforderung: Do

- In Ergänzung zu **[DS-R-PD-MA-08]** wird dokumentiert, inwiefern durch verteiltes Lernen der KI-Komponente das Ausspähen geschäftsrelevanter Information erschwert oder verhindert wird.

[DS-R-GI-MA-09] Signieren der Gewichte

Anforderung: Do

- Sofern gelernte Gewichte innerhalb der KI-Komponente ein schützenswertes Gut, etwa im Sinne des Urheberrechtes, darstellen, können diese mit einer digitalen Signatur[85], ähnlich einem Wasserzeichen, versehen werden. Dies ermöglicht, dauerhaft die Urheberschaft über das Modell nachzuweisen. Es ist zu dokumentieren, ob, in welchem Umfang und in welcher Form die Gewichte des ML-Modell signiert wurden.

9.2.3.3 Einbettung

[DS-R-GI-MA-10] Ungewollter Abfluss von Informationen

Anforderung: Do

- Falls gemäß **[DS-R-GI-RI-04]** das Risiko besteht, dass durch gezieltes Abfragen der KI-Anwendung geschäftsrelevante Informationen und insbesondere lizenzgebundene Daten ausgespäht werden, ist analog zu **[DS-R-PD-MA-09]** zu dokumentieren, welche Maßnahmen ergriffen wurden, um einen derartigen, ungewollten Abfluss von Informationen zu verhindern bzw. zu erschweren. Hierbei kann ggf. direkt auf **[DS-R-PD-MA-09]** verwiesen werden, oder auf ergriffene Maßnahmen zur Datenvorverarbeitung wie etwa **[DS-R-GI-MA-01]** bis **[DS-R-GI-MA-04]** sowie auf Dokumentationen aus dem **Risikogebiet: Integrität und Verfügbarkeit (IV)** der Dimension Sicherheit, beispielsweise die Beschränkung der Abfragemöglichkeiten (siehe **[SI-R-IV-MA-10]**).

[DS-R-GI-MA-11] Verhinderung von Modellextraktion

Anforderung: Do

- Es wird dargelegt, inwiefern die Ausgabe der KI-Anwendung ausschließlich die für deren Nutzung notwendigen Ergebnisse enthält/anzeigt.
  Beispiel: In der Regel ist es nicht erforderlich, dass die KI-Anwendung den gesamten *Softmax*-Vektor ausgibt, sondern es genügt für Nutzer*innen, zu wissen, welche Klasse den höchsten *Softmax*-Wert erreicht.
- Es wird dokumentiert, welche Informationen über die technischen Eigenschaften der KI-Anwendung bzw. ihrer KI-Komponente öffentlich zugänglich sind. Außerdem wird dargelegt, dass diese Informationen nicht über das notwendige Maß zur Information von Betroffenen bzw. Erklärbarkeit gegenüber Nutzer*innen (siehe **Dimension: Transparenz (TR)**) hinausgehen.
- Es wird dargelegt, inwiefern die gemäß **[DS-R-GI-RI-02]** sensiblen Modelleigenschaften angesichts der frei zugänglichen Informationen über die KI-Anwendung sowie der eingeräumten Abfragemöglichkeiten (siehe **[SI-R-IV-MA-10]**) vor Rekonstruktion geschützt sind. Unter Umständen werden mögliche Widersprüche bzw. Trade-Offs in Bezug auf die Information von Betroffenen sowie die Erklärbarkeit gegenüber Nutzer*innen (siehe **Dimension: Transparenz (TR)**) adressiert.

---

85 Siehe beispielsweise: Chen, H. et al., 2019. Deepmarks: A secure fingerprinting framework for digital rights management of deep learning models. In Proceedings of the 2019 on International Conference on Multimedia Retrieval. pp. 105–113. https://doi.org/10.1145/3323873.3325042 (letzter Aufruf: 23.06.2021)

9.2.3.4 Maßnahmen für den Betrieb

[DS-R-GI-MA-08] Speicherung und Löschung

Anforderung: Do

- In Ergänzung zu **[DS-R-PD-MA-10]** werden die technische Umsetzung und der Ort der Speicherung von lizenzgebundenen Daten bzw. geschäftsrelevanter Information dokumentiert. Dabei kann, falls möglich, direkt auf **[DS-R-PD-MA-10]** verwiesen werden, oder etwa auf Maßnahmen aus dem **Risikogebiet: Integrität und Verfügbarkeit (IV)** der Dimension Sicherheit, wie etwa **[SI-R-IV-MA-04]** und **[SI-R-IV-MA-05]**.
- Außerdem werden die eingesetzten technischen Verfahren zur Löschung von Daten bei Ablauf der Lizenz dokumentiert.
- Ferner wird erläutert, welche Maßnahmen ergriffen wurden, um die lizenzgebundenen Daten bzw. geschäftsrelevanten Informationen vor informationstechnischen Angriffen zu schützen. Hierbei kann u. a. auf Maßnahmen aus dem **Risikogebiet: Integrität und Verfügbarkeit (IV)**, beispielsweise **[SI-R-IV-MA-03]**, verwiesen werden.

9.2.4 Gesamtbewertung

[DS-R-GI-BW] Gesamtbewertung

Anforderung: Do

- Es liegt eine Dokumentation vor, in der abschließend und zusammenfassend begründet wird, dass das Restrisiko hinsichtlich der Vertraulichkeit geschäftsrelevanter Information im Kontext der KI-Anwendung gemäß **[DS-R-GI-KR-01]** vertretbar ist. Insbesondere wird begründet, dass lizenzgebundene Daten, die in der KI-Anwendung etwa zum Training oder als Input verarbeitet oder erhoben werden, nur gemäß der Lizenzbedingungen genutzt werden.
- Sofern nicht alle in **[DS-R-GI-KR-01]** spezifizierten Anforderungen erfüllt werden, sind die Abweichungen zu dokumentieren. Dies gilt ebenfalls für nur teilerfüllte Anforderungen, bei denen etwa die Kriterien nicht oder nicht immer erreicht wurden.

9.3 Risikogebiet: Beherrschung der Dynamik (BD)

Ziel des Risikogebiets Beherrschung der Dynamik ist es, sicherzustellen, dass der Datenschutz während des Betriebs der KI-Anwendung aufrechterhalten wird. Veränderte äußere Umstände können auch nach Inbetriebnahme der KI-Anwendung Maßnahmen bzw. Anpassungen erfordern. Beispielsweise können neue Technologien oder neu verfügbare Hintergrundinformationen das Risiko der Re-Identifizierbarkeit von Personen in einem Datensatz oder das Risiko der unerwünschten Einsehbarkeit sensibler Geschäftsdaten deutlich erhöhen. Außerdem kann eine Änderung der Rahmenbedingungen wie etwa der Gesetzeslage oder der Lizenzbedingungen Betreiber*innen und/oder Entwickler*innen vor neue Herausforderungen stellen. Nicht zuletzt nimmt auch das Verhalten von Nutzer*innen und Betroffenen Einfluss auf die Betriebsbedingungen der KI-Anwendung. So hängt beispielsweise die Einwilligung einer Person zur Verarbeitung ihrer Daten unter anderem davon ab, ob sie dem*der Betreiber*in vertraut, dass dieser*diese ihre Daten ausreichend schützen wird. Dieses Vertrauen kann durch äußere Umstände bestärkt oder andererseits durch öffentliche Skandale geschwächt werden.

9.3.1 Risikoanalyse und Zielvorgaben

[DS-R-BD-RI-01] Risikoanalyse und Zielvorgaben

Anforderung: Do

- **Risikoanalyse:** Es wird dokumentiert, ob und in welchem Umfang die KI-Anwendung während des Betriebs neu einkommende Daten verarbeitet, von welcher Kategorie (personenbezogen/geschäftsrelevant/lizenzgebunden) die neu einkommenden Daten sind und welche Anforderungen an sie gestellt werden. Außerdem werden die bestehenden Prozesse oder Mechanismen zur Kontrolle und zum Schutz der neu einkommenden Daten beschrieben. Es wird untersucht, welche Schäden entstehen, wenn die Anforderungen an die neu einkommenden Daten nicht erfüllt werden.
  Ferner wird analysiert, welche Rahmenbedingungen für die Datenverarbeitung durch die KI-Anwendung maßgeblich sind (z. B. Gesetze, interne Vorgaben) und an welchen äußeren Faktoren sich die Schutzmaßnahmen orientieren (z. B. verfügbares Hintergrundwissen, Wettbewerbssituation). Insbesondere wird abgeschätzt, mit welcher Wahrscheinlichkeit sich die Rahmenbedingungen oder relevante äußere Faktoren im Laufe des Betriebs ändern und welche Schäden daraus resultieren würden.
- **Zielvorgaben:** Basierend auf den ermittelten Risiken wird die Zielsetzung bezüglich des Datenschutzes während des Betriebs der KI-Anwendung formuliert. Diese enthält mindestens die Erfüllung der in **[DS-R-PD-KR-01]** und **[DS-R-GI-KR-01]** genannten Kriterien.

9.3.2 Kriterien zur Zielerreichung

[DS-R-BD-KR-01] Quantifizierung des Risikos

Anforderung: Do

- Um die Beherrschung der Betriebsdynamik angesichts neu einkommender Daten sowie möglicher Veränderungen der Rahmenbedingungen zu bewerten, werden Kriterien festgelegt und dokumentiert. Diese sollten mindestens die Einhaltung der Kriterien **[DS-R-PD-KR-01]** und **[DS-R-GI-KR-01]** beinhalten.
- Außerdem wird erläutert, dass die festgelegten Kriterien mit den Zielvorgaben in **[DS-R-BD-RI-01]** konform sind.

9.3.3 Maßnahmen

Die im Folgenden aufgelisteten Maßnahmen beziehen sich gleichermaßen auf Daten, KI-Komponente, Einbettung sowie den Betrieb.

9.3.3.1 Daten

9.3.3.2 KI-Komponente

9.3.3.3 Einbettung

9.3.3.4 Maßnahmen für den Betrieb

[DS-R-BD-MA-01] Einwilligung, Beschwerde, Löschung personenbezogener Daten

Anforderungen: Do | Pr

- Es liegt eine Dokumentation vor, die erläutert, wie im Betrieb der KI-Anwendung die datenschutzrechtlichen Vorgaben zur zweckgebundenen Einwilligung, Widerruf der Einwilligung, Beschwerde bei Verdacht auf Nichteinhaltung, sowie Löschung im Umgang mit personenbezogenen Daten umgesetzt werden. Wurden dazu betriebliche Prozesse etabliert, sind diese im Detail zu beschreiben.

[DS-R-BD-MA-02] Zukünftige Entwicklung bzgl. personenbezogener Daten

Anforderung: Do

- Es wird untersucht, abgeschätzt und dokumentiert, wie sich das Privacy-Risiko im Kontext der KI-Anwendung angesichts der Sammlung weiterer Daten (Trainings-, Input-, Nutzungsdaten) sowie im Hinblick auf zukünftig (allgemein) verfügbares Hintergrundwissen entwickeln wird.

[DS-R-BD-MA-03] Einwilligung, Beschwerde, Löschung lizenzgebundener Daten

Anforderungen: Do | Pr

- Es liegt eine Dokumentation vor, die erläutert, wie im Betrieb der KI-Anwendung die Einhaltung von Lizenzbedingungen im Umgang mit lizenzgebundenen Daten realisiert wird. Wurden dazu betriebliche Prozesse etabliert, sind diese im Detail zu beschreiben.

[DS-R-BD-MA-04] Zukünftige Entwicklung bzgl. geschäftsrelevanter Information

Anforderung: Do

- Es wird untersucht, abgeschätzt und dokumentiert, wie sich die kontextspezifische Bedeutung von geschäftsbezogener Information zukünftig entwickeln wird. Genauer gesagt, wird analysiert, ob absehbar ist, dass gewisse durch die KI-Anwendung verarbeitete Daten oder das Modell selbst zukünftig als geschäftsrelevante Information eingestuft werden bzw. diese Kategorisierung zukünftig nicht mehr für diese Daten zutrifft.

9.3.4 Gesamtbewertung

[DS-R-BD-BW] Gesamtbewertung

Anforderung: Do

- Unter Berücksichtigung der ergriffenen Maßnahmen wird dargelegt, dass das Restrisiko bezüglich des Datenschutzes im Betrieb der KI-Anwendung gemäß **[DS-R-BD-KR-01]** vertretbar ist.
- Sofern nicht alle in **[DS-R-BD-KR-01]** spezifizierten Anforderungen erfüllt werden, sind die Abweichungen zu dokumentieren. Dies gilt ebenfalls für nur teilerfüllte Anforderungen, bei denen etwa die Kriterien nicht, oder nicht immer, erreicht wurden.

Zusammenfassende Betrachtung

[DS-Z] Zusammenfassende Betrachtung der Dimension

Anforderung: Do

- Falls für diese Dimension ein mittlerer oder hoher Schutzbedarf besteht, ist eine Dokumentation über die verbleibenden Restrisiken zu erstellen. Zunächst werden die Restrisiken aus den verschiedenen Risikogebieten dieser Dimension zusammengefasst. Anschließend wird unter Berücksichtigung des Schutzbedarfs analysiert, ob die identifizierten Restrisiken insgesamt als vernachlässigbar, nicht vernachlässigbar (aber vertretbar) oder unvertretbar zu bewerten sind. Bei dieser Analyse sollten insbesondere die Auswirkungen von Maßnahmen aus der Dimension Sicherheit berücksichtigt werden, falls diese dazu beitragen, die Risiken hinsichtlich des Datenschutzes abzuschwächen oder zu beherrschen. Das Ergebnis der Analyse ist zu erläutern.
- Falls potenziell negative Auswirkungen von Risiken oder Maßnahmen dieser Dimension auf andere Dimensionen wie etwa Transparenz festgestellt wurden, sind diese zu dokumentieren.
- Es wird ein Fazit über die Dimension gezogen, welches insbesondere die Bewertung der Restrisiken enthält.

10. Dimensionsübergreifende Beurteilung der Vertrauenswürdigkeit (BV)

Die Diskussion der einzelnen Risikogebiete in den vorhergehenden Kapiteln umfasst zum Schluss stets eine Würdigung der ergriffenen Maßnahmen (Gesamtbewertung), welche argumentiert, dass diese Maßnahmen ausreichen, um die basierend auf der Risikoanalyse definierten Qualitätskriterien zu erfüllen. Wie in den zusammenfassenden Betrachtungen der einzelnen Dimensionen dargelegt, kann es allerdings Zielkonflikte zwischen unterschiedlichen Qualitätsdimensionen geben. Ziel dieses Kapitels ist es, darzustellen, wie mit solchen Zielkonflikten umzugehen ist.

Zielkonflikte zwischen den Dimensionen können etwa aus mangelnder Realisierbarkeit entgegengesetzter Anforderungen resultieren bzw. daraus, dass die Erfüllung von Anforderungen der einen Dimension die Risiken hinsichtlich einer anderen Dimension steigern würde. Die Komplexität möglicher Qualitätsanforderungen lässt sich eingehend am Beispiel einer KI-Anwendung zur Bewertung der Kreditwürdigkeit verdeutlichen. Bereits bei der Auswahl der Features, d. h. der in den Eingabedaten enthaltenen Merkmale, auf denen die KI-Komponente operiert, muss eine Abwägung stattfinden. Denn Anforderungen hinsichtlich Datensparsamkeit und Fairness, da sensible Persönlichkeitsmerkmale bei der Entscheidung über die Kreditwürdigkeit keinen Einfluss haben sollten, könnten in Konflikt zu dem Ziel einer hohen *Accuracy* stehen, die meist durch das Bereitstellen möglichst vieler Informationen (Features) über eine betreffende Person gesteigert wird. Ein weiterer Zielkonflikt hinsichtlich der Performanz könnte sich zudem aus der Wahl eines Fairness-Konzepts ergeben. Beispielsweise steht die Umsetzung von Gruppenfairness angesichts einer »unfairen« Datenbasis einer perfekten Prädiktion (verglichen mit dem unfairen Datensatz) entgegen. Auch bei der Wahl des Modells kommen in der Regel Zielkonflikte auf. So können Modelle, die zuverlässige Ergebnisse über die Kreditwürdigkeit potenzieller Kund*innen liefern, ggf. für Expert*innen nicht interpretierbar sein. In diesem sensiblen Einsatzkontext könnten, nach sorgfältiger Abwägung, Einbußen an Performanz hingenommen werden, sofern dafür ein interpretierbares Modell zum Einsatz kommt. Nicht zuletzt sind in diesem Beispiel auch die Eingriffsmöglichkeiten im Betrieb sowie der Umfang an Informationen, die beispielsweise Mitarbeitenden oder Kund*innen über das System bereitgestellt werden, zu diskutieren. Hier steht das essenzielle Gebot der menschlichen Aufsicht und Autonomie unter Umständen in Konflikt mit der Sicherheit in dem Sinne, dass Möglichkeiten zum Angriff oder zur Manipulation der KI-Anwendung eröffnet bzw. erleichtert werden könnten.

Um eine nachhaltige Abwägung der bestehenden Zielkonflikte sowie der damit verbundenen Restrisiken zu ermöglichen, ist es wichtig, alle wesentlichen Stakeholder-Interessen zu berücksichtigen. Insbesondere sollten Risiken und deren Effekte unter zwei verschiedenen Gesichtspunkten, wie nachfolgend dargestellt, betrachtet werden. In den sechs Dimensionen werden Risiken in erster Linie hinsichtlich potenzieller Auswirkungen auf Nutzer*innen, Betroffene oder die (unmittelbare) Umgebung untersucht. Gleichzeitig haben Risiken wie etwa fehlerhaftes oder gar schädliches Verhalten einer KI-Anwendung auch Auswirkungen auf die sie betreibende Organisation. So besteht bei einer KI-Anwendung zur Kreditvergabe zum Beispiel das Risiko der Diskriminierung, wobei Persönlichkeitsrechte von Kund*innen verletzt werden, womit jedoch auch eine Rufschädigung des betreffenden Kreditinstituts einhergeht. Dieses Beispiel zeigt, dass KI-Risiken in der Entscheidungsbildung einer Organisation berücksichtigt werden müssen. Organisationen, die KI-Anwendungen einsetzen oder betreiben,

sollten hierzu eine KI-Governance etablieren sowie Organisationsstrukturen[86], die Rollen und Verantwortlichkeiten bezüglich des KI-(Risiko-)Managements regeln. In der Umsetzung schließt dies einen entsprechenden Prozess mit ein, um Zielkonflikte und potenzielle Restrisiken abzuwägen. Insbesondere wird innerhalb einer Organisation eine Instanz benötigt, die das Ergebnis des Abwägungsprozesses bestätigt und die Verantwortung für die damit verbundenen Restrisiken trägt.

Hierzu schlägt die High-Level Expert Group on AI etwa ein »AI Ethics Review Board«[87] vor, das die Verantwortlichkeiten und die ethische Praxis angesichts des Einsatzes von KI diskutiert, sowie Prozesse zur kontinuierlichen Bewertung des Systems.

Aus der vorangegangenen Diskussion ergibt sich somit folgende Anforderung:

[BV] Dimensionsübergreifende Beurteilung der Vertrauenswürdigkeit der KI-Anwendung

Anforderung: Do

- Sofern in einer Dimension mit mittlerem oder hohem Schutzbedarf das Fazit gezogen wurde, dass unvertretbare Restrisiken bestehen, so ist die KI-Anwendung nicht als vertrauenswürdig einzustufen.
- Wurden zwar keine unvertretbaren, aber dennoch nicht vernachlässigbare Restrisiken identifiziert, so ist zu untersuchen, inwiefern diese mit potenziellen Zielkonflikten zwischen den Dimensionen zusammenhängen. Dazu wird insbesondere erörtert, inwieweit Restrisiken in einer Dimension ggf. unvermeidbar sind, um Risiken in einer anderen Dimension zu mindern. Falls argumentiert wird, dass ein Restrisiko aufgrund eines Zielkonflikts nicht abgeschwächt werden kann bzw. soll, so ist die gewählte Priorisierung in Bezug auf den vorliegenden Trade-Off abzuwägen und zu begründen. Bei der Begründung sollte insbesondere der Schutzbedarf der betrachteten Dimensionen berücksichtigt werden.
  - Falls nicht plausibel begründet werden kann, dass die vorhandenen Restrisiken aufgrund der Existenz von Zielkonflikten unvermeidbar sind, ist die KI-Anwendung nicht als vertrauenswürdig einzustufen.
  - Falls plausibel dargelegt werden kann, dass alle bestehenden Restrisiken aufgrund kaum vermeidbarer Zielkonflikte in Kauf genommen werden müssen, und die Priorisierung in Bezug auf die vorhandenen Trade-Offs erläutert wird, so ist es möglich, die KI-Anwendung trotz nicht vernachlässigbarer Restrisiken als vertrauenswürdig zu beurteilen. Die Beurteilung, ob die KI-Anwendung vertrauenswürdig ist, ist ausführlich zu erläutern.
- Wurde in jeder Dimension mit mittlerem oder hohem Schutzbedarf das Fazit gezogen, dass die Restrisiken vernachlässigbar sind, so ist die KI-Anwendung als vertrauenswürdig zu beurteilen.

86 Für eine ausführliche Betrachtung solcher Fragestellungen siehe die Fraunhofer IAIS Studie »AI Management Systems« (in Erscheinung), die Anforderungen an Organisationen zum Umgang mit KI in Bezug auf Governance, Management und technisch-organisatorische Maßnahmen diskutiert und dabei die aktuellen Standardisierungsaktivitäten des ISO/IEC JTC1/SC 42 »Artificial Intelligence« miteinbezieht.

87 High-Level Expert Group on AI (HLEG) (Juli 2020). The Assessment List for Trustworthy Artificial Intelligence (ALTAI). Veröffentlicht von der Europäischen Kommission. https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment (letzter Aufruf: 21.06.2021)

Impressum

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

Redaktion
Daria Tomala
Silke Loh

Grafik und Layout
Achim Kapusta

Bildquellen
Titelbild: Alex – stock.adobe.com
S. 8, Foto Prof. Dr. Andreas Pinkwart: ©MWIDE NRW/E. Lichtenscheid

Stand
Juli 2021

1. Auflage

Kontakt

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

www.iais.fraunhofer.de

Ansprechpartner:
Dr. rer. nat. Maximilian Poretschkin
maximilian.poretschkin@iais.fraunhofer.de