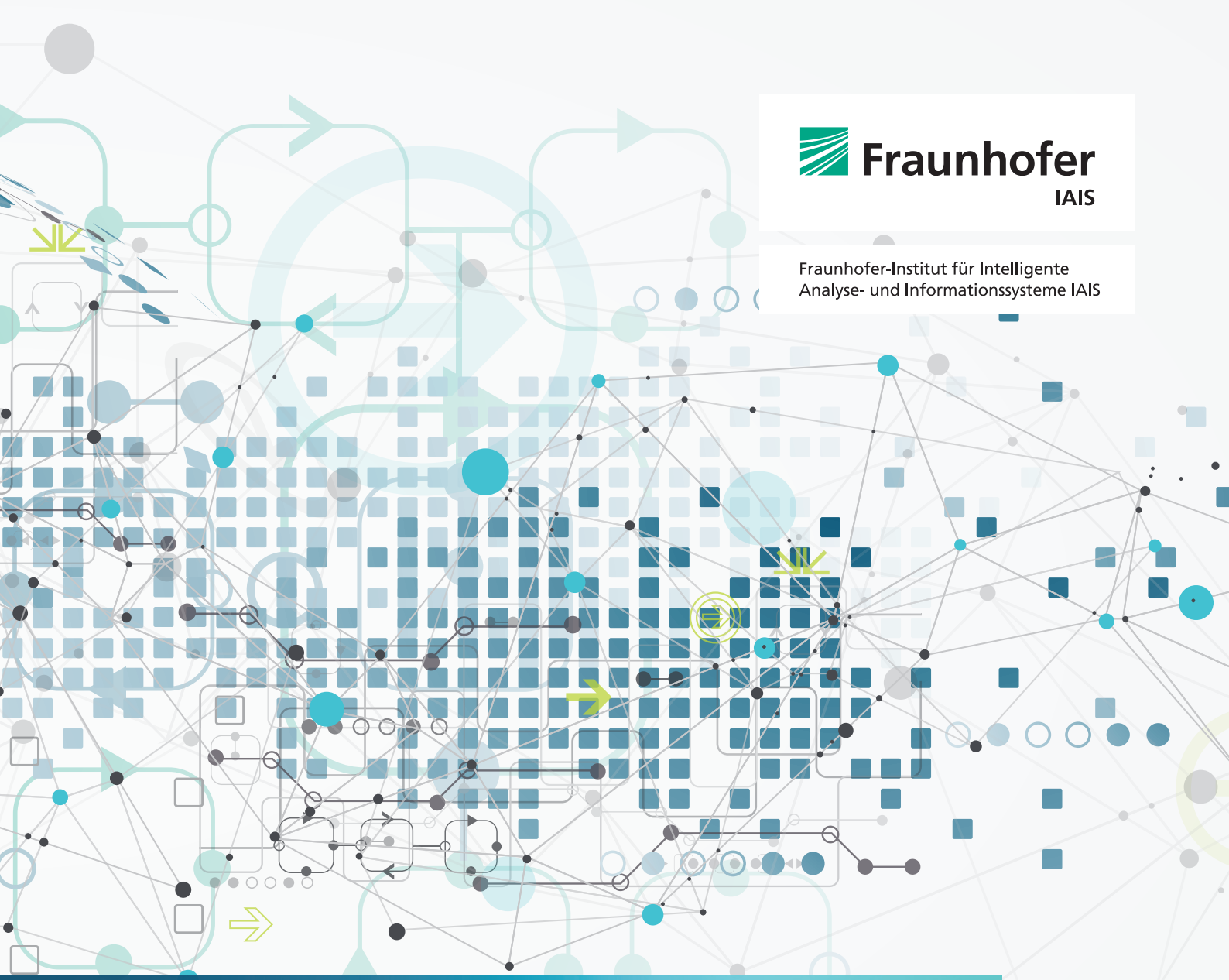


Fraunhofer IAIS

Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS

Whitepaper

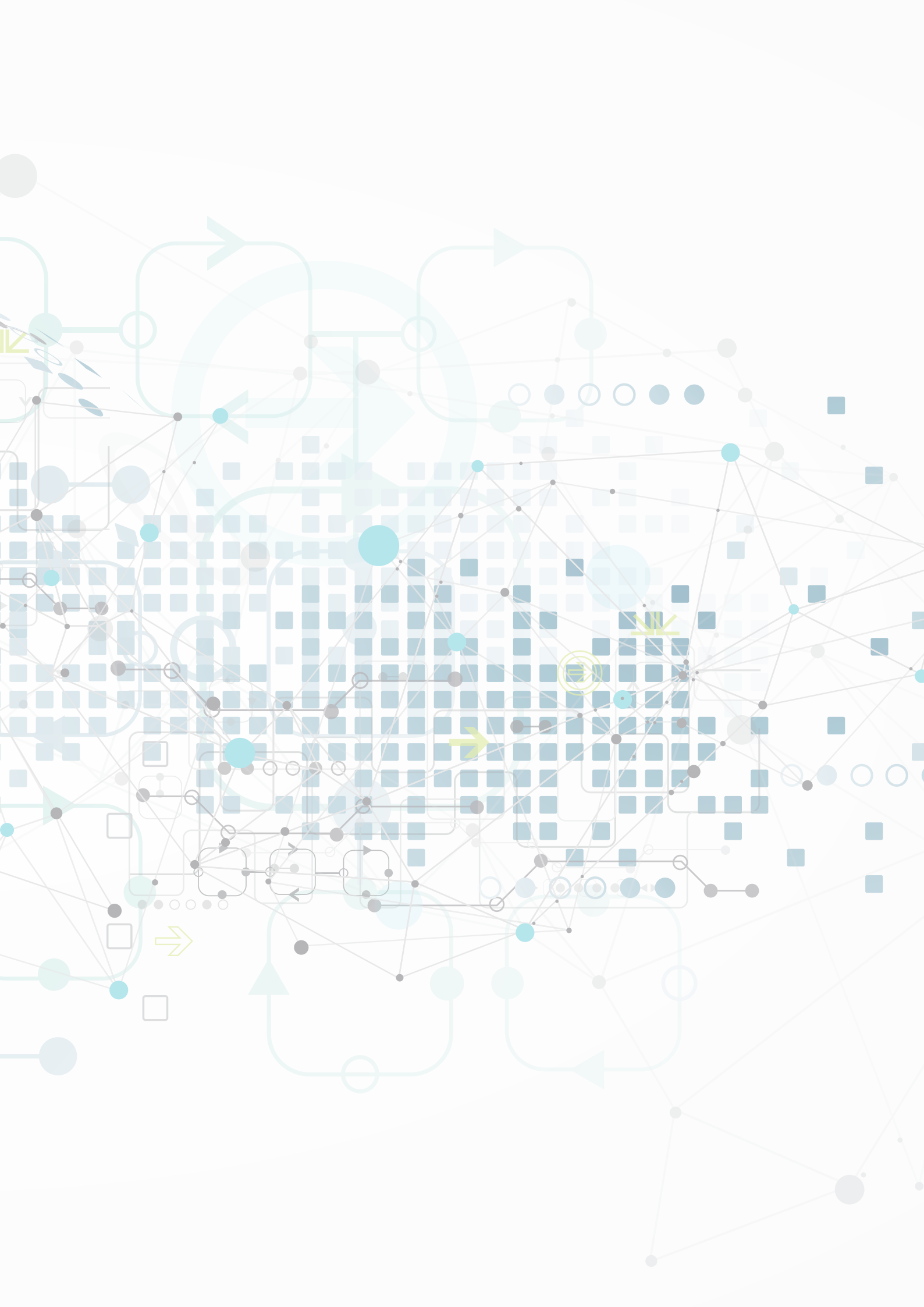
Vertrauenswürdige KI-Anwendungen mit Foundation-Modellen entwickeln

POWERED BY KI NRW

Gefördert durch:
Ministerium für Wirtschaft, Industrie, Klimaschutz und Energie des Landes Nordrhein-Westfalen

ZERTIFIZIERTE KI
Qualität sichern. Fortschritt gestalten.

UNIVERSITÄT BONN



Fraunhofer IAIS

Whitepaper

Vertrauenswürdige KI-Anwendungen mit Foundation-Modellen entwickeln

Autorinnen und Autoren
Michael Mock [1]
Sebastian Schmidt [1]
Felix Müller [2, 1]
Rebekka Görge [1]
Anna Schmitz [1]
Elena Haedecke [2, 1]
Angelika Voss [1]
Dirk Hecker [1]
Maximilian Poretschkin [1, 2]

[1] Fraunhofer IAIS, [2] Universität Bonn

Januar 2024

www.iais.fraunhofer.de/zertifizierte-ki

KI.NRW

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

Das Fraunhofer IAIS

Als Teil der größten Organisation für anwendungsorientierte Forschung in Europa ist das Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS mit Sitz in Sankt Augustin/Bonn und einem Standort in Dresden eines der führenden Wissenschaftsinstitute auf den Gebieten Künstliche Intelligenz (KI), Maschinelles Lernen und Big Data in Deutschland und Europa.

Rund 350 Mitarbeitende unterstützen Unternehmen bei der Optimierung von Produkten, Dienstleistungen und Prozessen sowie bei der Entwicklung neuer digitaler Geschäftsmodelle. Das Fraunhofer IAIS gestaltet die digitale Transformation unserer Arbeits- und Lebenswelt: mit innovativen KI-Anwendungen für Industrie, Gesundheit und Nachhaltigkeit, mit zukunftsweisenden Technologien wie großen KI-Sprachmodellen oder Quantum Machine Learning, mit Angeboten für die Aus- und Weiterbildung oder für die Prüfung von KI-Anwendungen auf Sicherheit und Vertrauenswürdigkeit.

www.iais.fraunhofer.de

Inhalt

Abbildungsverzeichnis

Executive Summary

Die Vertrauenswürdigkeit von KI-Anwendungen ist seit einiger Zeit Gegenstand der Forschung und wird auch mit der geplanten KI-Verordnung der EU adressiert. Mit den aktuell aufkommenden Foundation-Modellen im Bereich der Text-, Sprach- und Bildverarbeitung bieten sich völlig neue Möglichkeiten, KI-Anwendungen zu entwickeln. Dieses Whitepaper zeigt auf, wie die Vertrauenswürdigkeit einer mit Foundation-Modellen entwickelten KI-Anwendung bewertet und sichergestellt werden kann. Zu diesem Zweck wird die anwendungsspezifische, risikobasierte Vorgehensweise zur Prüfung und Sicherstellung der Vertrauenswürdigkeit von KI-Anwendungen, wie sie im »KI-Prüfkatalog zur Gestaltung vertrauenswürdiger Künstlicher Intelligenz« des Fraunhofer IAIS entwickelt worden ist, in den Kontext von Foundation-Modellen übertragen. Dabei wird besonders berücksichtigt, dass sich spezielle Risiken der Foundation-Modelle auf die KI-Anwendung auswirken können und zusätzlich bei der Prüfung der Vertrauenswürdigkeit beachtet werden müssen.

Kapitel 1 des Whitepapers erläutert den grundsätzlichen Zusammenhang zwischen Foundation-Modellen und auf ihnen basierenden KI-Anwendungen in Bezug auf die Vertrauenswürdigkeit. Kapitel 2 gibt eine Einführung in die technische Konstruktion von Foundation-Modellen und Kapitel 3 zeigt, wie darauf aufbauend KI-Anwendungen entwickelt werden können. Kapitel 4 gibt einen Überblick über die dabei entstehenden Risiken in Hinblick auf die Vertrauenswürdigkeit. Kapitel 5 zeigt, welche Anforderungen an KI-Anwendungen und Foundation-Modelle aus der bevorstehenden KI-Verordnung der Europäischen Union zur erwarten sind und Kapitel 6 zeigt schließlich die Systematik und Vorgehensweise, um Anforderungen an die Vertrauenswürdigkeit zu erfüllen.

1 Einführung

Mit dem Erscheinen von Foundation-Modellen ist eine neue Zeitrechnung der Entwicklung von Künstlicher Intelligenz (KI) eingeleitet worden. Diese Modelle werden auf großen multimodalen Datenmengen trainiert und können zum Beispiel per Fine-Tuning an eine große Anzahl an Aufgaben angepasst werden. Sie unterscheiden sich damit fundamental von »herkömmlichen« Machine Learning (ML) Anwendungen, welche üblicherweise für eine bestimmte Aufgabe trainiert werden. Die Fähigkeiten und Funktionalitäten von Foundation-Modellen verändern somit drastisch die Art und Weise, wie KI-Anwendungen entwickelt und eingesetzt werden. Insbesondere ermöglichen sie eine Vielzahl von Anwendungen, die bislang mit herkömmlichen Methoden nur schwer realisierbar schienen, etwa realistisch wirkende Unterhaltungen zu komplexen Themen mit einem Chatbot über eine längere Zeit hinweg – mithin also das Bestehen des Turing-Tests –, oder eine detaillierte und sogar interpretierende Beschreibung von anspruchsvollen Bildinhalten. Besonders an den Foundation-Modellen ist zudem, dass sie ohne viel Aufwand an die unterschiedlichsten Aufgaben angepasst werden können bzw. ohne Anpassung direkt genutzt werden können. So kann jedermann ohne KI-Vorkenntnisse durch Prompting von ChatGPT etwa Vorlagen für Redemanuskripte, E-Mails oder Projektberichte erstellen lassen. Während bis zum Erscheinen von Foundation-Modellen vor allem repetitive Aufgaben, beispielsweise im Rahmen von Prozessautomatisierungen, durch KI-Anwendungen beeinflusst und sogar vollkommen übernommen wurden, betrifft dies nun auch kreative Berufsbilder, zum Beispiel in der Mediengestaltung. Das sich hieraus ergebende wirtschaftliche Potenzial ist gewaltig und wird auf 2,4 bis 4,4 Billionen US Dollar [CH23a] geschätzt. In der Konsequenz unterliegen KI-Liefer- und Wertschöpfungsketten gewaltigen Veränderungen. Gleichzeitig werden weltweit in allen größeren Wirtschaftsräumen, allen voran der Europäischen Union, Regulierungsanstrengungen für Künstliche Intelligenz unternommen.[1] Ein wesentlicher Baustein der Europäischen KI-Verordnung (EU AI Act) [EU23] ist eine Konformitätsbewertung von Hochrisikosystemen vor dem Inverkehrbringen. Eine notwendige Voraussetzung für die Durchführung solcher KI-Konformitätsbewertungen stellen KI-spezifische Prüfverfahren dar. Eine Herausforderung hierbei ist, dass KI-Anwendungen, die durch Maschinelles Lernen realisiert werden, nicht explizit programmiert sind, sondern die Funktionalität aus großen Datenbeständen erlernt wird. In der Folge greifen herkömmliche Verfahren zur Qualitätssicherung und zum Qualitätsnachweis von Software zu kurz. Ein wichtiger Ansatz an dieser Stelle besteht darin, das aus der Automobilbranche bekannte Konzept einer Operational Design Domain auf andere Kontexte und Domänen dahingehend zu erweitern, dass der Eingaberaum mit einer semantischen Struktur versehen wird, welche eine systematische Suche nach Schwachstellen der KI-Anwendung erlaubt und als Grundlage für strukturierte Absicherungsargumentationen und Safety Cases dienen kann. Eine andere Herausforderung besteht darin, dass KI-Anwendungen allgemein komplexen Wertschöpfungsketten unterliegen. Wichtige Teilnehmende der Wertschöpfungskette sind etwa Hardwareanbieter, Datenanbieter, Frameworkprovider sowie Anbieter von KI-Basisdiensten (etwa OCRs) und -Toolboxen. Die Frage nach den Übergabepunkten der Verantwortung für einzelne Komponenten einer KI-Anwendung entlang der Wertschöpfungskette – etwa wie eine nicht zugängliche OCR als Teil einer zu prüfenden KI-Anwendung zu bewerten ist – ist aktuell in der Praxis oftmals noch eine Herausforderung. Gleichzeitig sind in der jüngeren Vergangenheit eine Reihe von vielversprechenden Ansätzen zur Prüfung von KI-Anwendungen veröffentlicht worden [PO21, LE21, NI23, AI23, MO23], welche bereits Gegenstand verschiedener Standardisierungsaktivitäten sind. Zudem lässt sich aktuell die Bildung einer KI-Prüfindustrie beobachten.

Für den Fall der Prüfung von KI-Anwendungen, welche auf Foundation-Modellen basieren oder diese als Komponente inkludieren, verschärfen sich die eben beschriebenen Herausforderungen noch einmal: Dies liegt einerseits daran, dass es praktisch unmöglich ist, ein Analogon zu dem aus der Automotive-Domäne bekannten Konzept der Operational Design Domain zu konstruieren, also eine systematische Beschreibung des möglichen Eingaberaums für Foundation-Modelle. Anstelle von gezielten Tests treten somit Benchmarks, welche die Performanz unterschiedlicher Foundation-Modelle für bestimmte Aufgabentypen systematisch vergleichen. Weiterhin beziehen herkömmliche Prüfverfahren den Einsatzkontext der KI-Anwendung mit ein, etwa für eine Risikoanalyse möglicher, durch das System potenziell zu erwartender Schäden. Angesichts der mannigfaltigen Einsatzmöglichkeiten von Foundation-Modellen ist ein solcher Einsatzkontext aber ebenfalls schwierig zu definieren, was im Rahmen der Finalisierung des Entwurfs der KI-Verordnung zu der Diskussion geführt hat, ob Foundation-Modelle somit grundsätzlich als Hochrisikosysteme einzustufen sind. Schließlich verschärfen Foundation-Modelle noch einmal die Wertschöpfungskettenproblematik, da die Inklusion eines solchen Foundation-Modells eine ganz andere Komplexität mit sich bringt, als der Einsatz einer OCR als Drittkomponente. Gleichzeitig können gerade auf Foundation-Modellen besonders viele KI-Dienste

1 Zum Beispiel wurde im Oktober 2023 ein »Code of Conduct« für Künstliche Intelligenz durch die G7 veröffentlicht [BMDV23].

aufgesetzt werden. Eine Prüfung von auf Foundation-Modellen basierenden KI-Anwendungen muss daher zwei Blickwinkel einnehmen: Zum einen die Frage der Korrektheit und Vertrauenswürdigkeit des Foundation-Modells selbst, zum anderen die Frage, ob ein hierauf aufgesetzter KI-Service oder ein hierauf aufgesetztes System die geforderte Downstream-Task korrekt implementiert. Angesichts der Marktmacht der Anbieter solcher Foundation-Modelle setzt eine Antwort auf die erste Frage einen entsprechenden Druck regulatorischer oder marktwirtschaftlicher Vorgaben voraus. Besonders drängend ist die zweite Frage, nämlich die nach der Korrektheit von Services oder Systemen, die auf Foundation-Modellen aufsetzen. Denn solche Services und Systeme sind so schnell und einfach zu implementieren, dass in der Praxis nicht immer ein flankierender Qualitätssicherungsprozess zu erwarten ist. Das wiederum erhöht die Eintrittswahrscheinlichkeit von Fehlern.

Das vorliegende Whitepaper setzt hier an und zeigt einen systematischen Zugang zur Prüfung von KI-Anwendungen, die mit Foundation-Modellen realisiert werden. Es stellt eine risikobasierte Systematik vor, mit der die Vertrauenswürdigkeit solcher KI-Anwendungen bewertet und sichergestellt werden kann. Dazu gibt es einen Überblick über die sich aus der Nutzung von Foundation-Modellen ergebenden Risiken, zeigt aber auch neue Möglichkeiten (zum Beispiel bei der Entwicklung von Tests) für die Entwicklung vertrauenswürdiger KI-Anwendungen auf.

Um den Zusammenhang zwischen KI-Anwendungen, Vertrauenswürdigkeit und Foundation-Modellen zu motivieren, stellt Abbildung 1 zunächst einen Bezug zur klassischen Softwareentwicklung her und betrachtet den Spezialfall, dass ein Foundation-Modell einen Software-Code erzeugt. Diese Funktionalität wird zurzeit bereits als Assistenzfunktion angeboten, wie zum Beispiel durch den Github Co-Pilot [NG22], und könnte in Zukunft auch weiter automatisiert werden.

In der Softwareentwicklung stellt die testbasierte Entwicklung den State of the Art dar. Softwaretests überprüfen an einer Menge von diversen Testfällen, ob die Anwendung gemäß ihrer Spezifikation funktioniert. Im Rahmen der sogenannten agilen Softwareentwicklung werden diese Tests sogar kontinuierlich während der Softwareentwicklung durchgeführt und

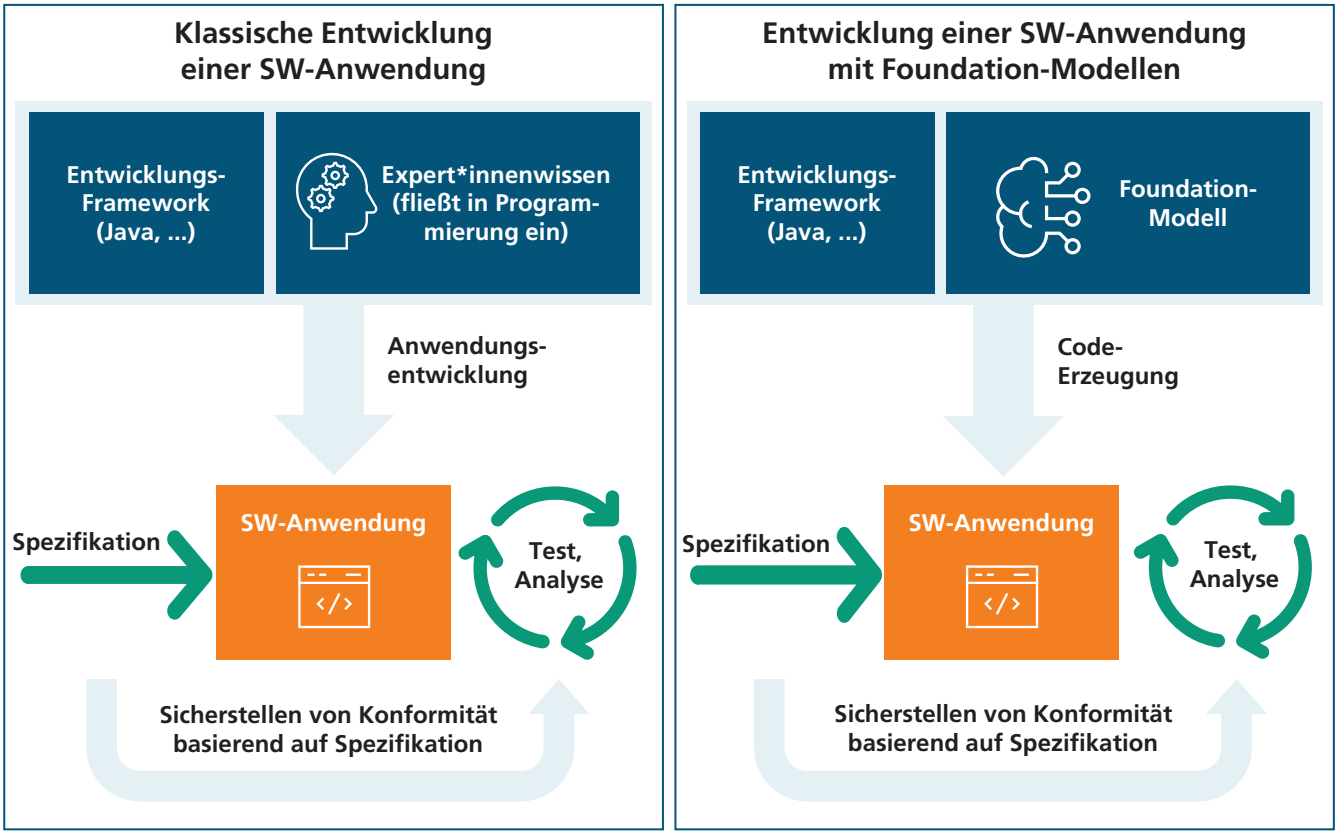


Abbildung 1: Spezifikation und Test in der Entwicklung einer Software-Anwendung (SW-Anwendung) erfolgen immer in Bezug auf den bestimmten Anwendungsfall. Das Vertrauen in die Konformität der Anwendung wird durch die klassischen Software-Testverfahren erreicht, welche auf der Spezifikation basierend durchgeführt werden. Links ist der Prozess zur klassischen Entwicklung einer SW-Anwendung dargestellt, während rechts ein Foundation-Modell zur Code-Erzeugung der SW-Anwendung genutzt wird.

unterstützen dabei, hochqualitative Software auch in großen Entwicklerteams schnell gemeinsam zu erstellen.

Würde man nun in einem solchen Team nicht mehr jede einzelne Zeile Code von einem Entwickler oder einer Entwicklerin schreiben lassen, sondern auch Code[2] einsetzen wollen, der von einem Foundation-Modell erzeugt wird, so gäbe es keinen Anlass, auf die Tests zu verzichten. Vielmehr würde gerade das Vorhandensein der Tests aufzeigen, dass es möglich ist, auch den durch ein Foundation-Modell generierten Code zu nutzen. Wie im Fall der klassischen Softwareentwicklung würden die Tests somit das Vertrauen in die Funktionsfähigkeit der Anwendung rechtfertigen.

Diese Situation lässt sich auch auf den Fall der Entwicklung von KI-Anwendungen übertragen, wie in Abbildung 2 skizziert.

KI-Anwendungen unterliegen bekanntermaßen bestimmten KI-spezifischen Risiken, welche verschiedene Ziele, Stakeholder und Eigenschaften der Anwendung betreffen können und mitunter als Dimensionen der Vertrauenswürdigkeit bezeichnet werden.[3] Die Vertrauenswürdigkeit einer KI-Anwendung ist insbesondere mit der Beherrschung dieser Risiken assoziiert und Anforderungen an die Vertrauenswürdigkeit erstrecken sich über die verschiedenen Dimensionen hinweg, zum Beispiel Anforderungen an hinreichende Zuverlässigkeit oder Fairness. Da die Anwendung nun nicht mehr nach einer vom Menschen vorgegebenen Berechnungsvorschrift, sondern mit einem »trainierten Modell« arbeitet, stellt sich insbesondere die Frage, ob sie auch auf den realen Anwendungsbereich generalisiert, also tatsächlich auch auf neuen Eingabedaten funktioniert, die nicht in den Trainingsdaten vorhanden waren.

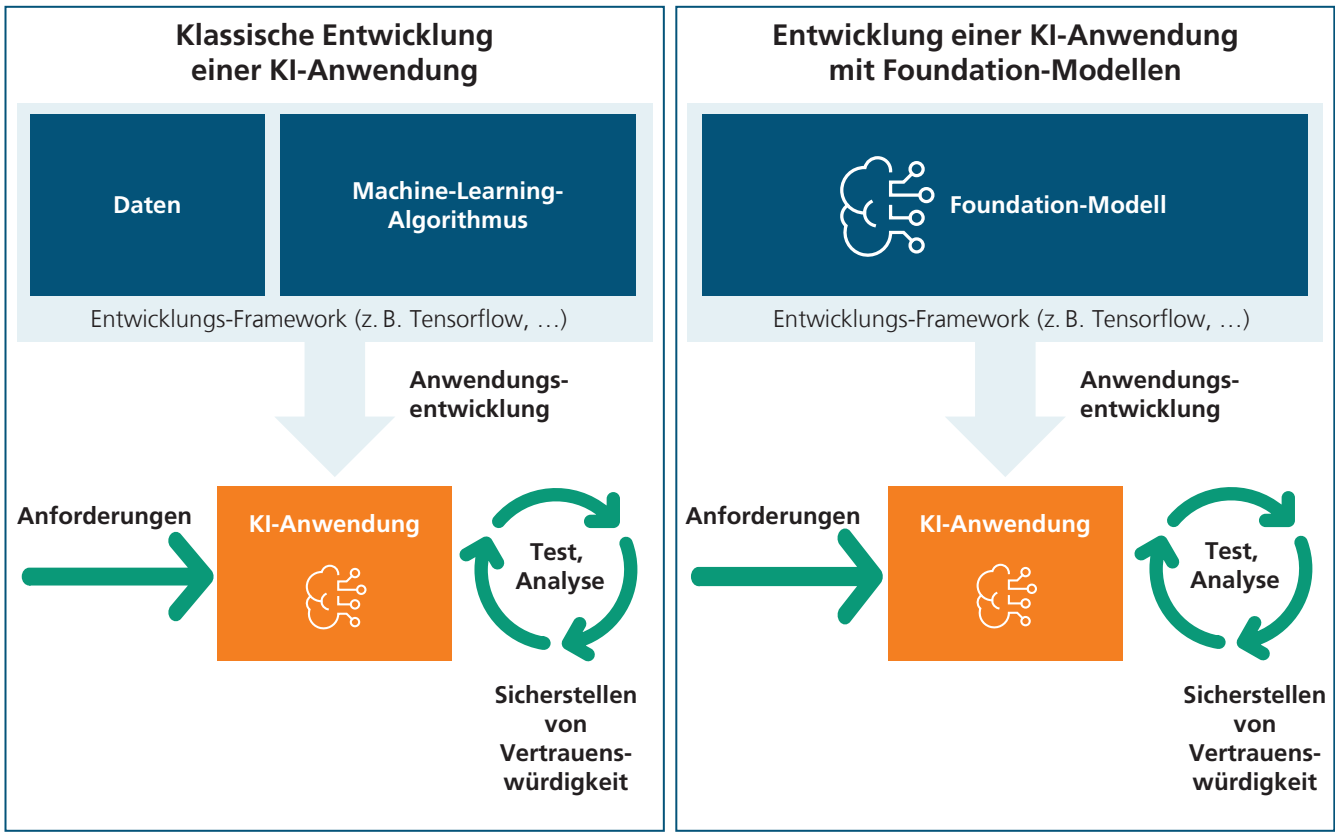


Abbildung 2: Die Anforderungen an die Vertrauenswürdigkeit und ihre Überprüfung geschieht bei der Entwicklung von KI-Anwendungen auf der Anwendungsebene. Der linke Teil der Abbildung zeigt den klassischen Entwicklungsprozess im Vergleich zum rechten Teil, wo Foundation-Modelle zur Entwicklung genutzt werden.

2 Ein Beispiel hierfür ist die KI-basierte Anwendung »GitHub-Copilot«, welcher Programmierer*innen durch Generieren und Autovervollständigen von Code unterstützt [GHC23].

3 Der »KI-Prüfkatalog« des Fraunhofer IAIS unterscheidet zwischen sechs Risikodimensionen der Vertrauenswürdigkeit [PO21].

Wie hoch diese Anforderungen in den verschiedenen Dimensionen sind, hängt vom Anwendungsfall ob. So wird eine sicherheitskritische Anwendung, wie zum Beispiel die Personenerkennung im automatisierten Fahren, höhere Anforderungen an die Zuverlässigkeit erfüllen müssen als eine KI-basierte Software, die Kaufempfehlungen in einem Web-Shop gibt. Manche KI-Anwendungen unterliegen hohen Anforderungen an Fairness, zum Beispiel bei der automatisierten Vergabe von Krediten, andere Anwendungen haben bezüglich Fairness typischerweise geringere Anforderungen, wie etwa bei der Qualitätssicherung in automatisierten Fertigungsprozessen. Erschwerend kommt hinzu, dass es Trade-offs zwischen den verschiedenen Dimensionen der Vertrauenswürdigkeit gibt. Die Abwägung von verschiedenen Anforderungen richtet sich nach dem Zweck der KI-Anwendung und resultiert nicht aus der zur Entwicklung der Anwendung genutzten Technologie. Auch der aktuelle Entwurf der EU-Verordnung zur KI [EU23] klassifiziert KI-Anwendungen in Risikostufen je nach Anwendungszweck, und nicht etwa nach Technologiestufen.

Für den Nachweis der anwendungsspezifischen Anforderungen kommen bei KI-Anwendungen, wie in der klassischen Softwareentwicklung, Tests zum Einsatz. Diese Tests messen, je nach Anforderungsprofil, nicht allein die Verlässlichkeit, sondern auch weitere Eigenschaften der Anwendung. Eine zentrale Rolle hierbei spielen die Testdaten. Sie sollten den Anwendungsbereich bestmöglich abdecken. Eine korrekte Funktion der KI-Anwendung auf den Testdaten schafft wieder Vertrauen in die Funktionsfähigkeit der realen KI-Anwendung.

Die Foundation-Modelle bieten nun eine völlig neue Grundlage, um KI-Anwendungen zu entwickeln. Foundation-Modelle können zum Beispiel mit deutlich weniger Daten für spezialisierte Anwendungen nachtrainiert werden (sog. Fine-Tuning), als es für eine Neuentwicklung der Anwendung erforderlich wäre. Darüber hinaus gibt es auch neue Verfahren, KI-Anwendungen zu entwickeln (siehe Kapitel 3), die Foundation-Modelle gewissermaßen als »Library« nutzen und aufrufen. Jedoch gilt weiterhin, wie in der klassischen Softwareentwicklung, dass sowohl die Anforderungen an die Vertrauenswürdigkeit der KI-Anwendung als auch ihr Nachweis anwendungsspezifisch sind und insbesondere anwendungsspezifische Tests (und Testdaten) erfordern. Neben den neuen Möglichkeiten, KI-Anwendungen zu entwickeln, eröffnen Foundation-Modelle auch neue Perspektiven, KI-Anwendungen zu testen, indem man Testdaten mit Foundation-Modellen erzeugt.

Dieses Whitepaper gibt einen Einblick in die Zusammenhänge und zeigt einen praktikablen Weg zur Entwicklung vertrauenswürdiger KI-Anwendungen mit Foundation-Modellen.

2 Die Grundlagen von Foundation-Modellen

Was macht nun die Foundation-Modelle aus technischer Sicht zu einem so mächtigen Baustein und wie unterscheiden sie sich von bisherigen KI-Modellen?

Grundsätzlich werden alle KI-Modelle auf Daten trainiert, um das in den Daten vorhandene immanente Wissen algorithmisch nutzbar zu machen. Man unterscheidet dabei klassisch zwischen dem sogenannten »überwachten (supervised) Lernen« und dem »unüberwachten (unsupervised) Lernen«, wobei mit »Lernen« in beiden Fällen ein auf stochastischen Gesetzmäßigkeiten beruhender Optimierungsprozess gemeint ist (sogenanntes Maschinelles Lernen). Die beiden Varianten sind schematisch in Abbildung 3 dargestellt.

Im Fall des überwachten Lernens aus Abbildung 3 wird ein Modell $\hat{f}$ trainiert, um möglichst gut eine vorgegebene Zielfunktion f zu approximieren. Die Wirkweise der Zielfunktion f ist dabei unbekannt, ansonsten hätte sie auch als klassischer Algorithmus programmiert werden können. Stattdessen wird die Zielfunktion f nur über Beispiele (Trainingsdaten) spezifiziert. Ein Modell zur Spracherkennung würde etwa mit Tonaufnahmen von gesprochenen Beispielsätzen trainiert, zu denen auch die zugehörigen Texte als sogenannte »Labels« vorliegen und zusammen die sogenannten Trainingsdaten bilden. Der Lernalgorithmus passt das Modell $\hat{f}$ iterativ an die Trainingsdaten an, sodass der gesamte Fehler $|f\text{-}\hat{f}|$ auf den Trainingsdaten möglichst klein wird. Abstrakt gesehen wird dabei das bislang nur immanent vorhandene Wissen über den Wirkzusammenhang zwischen den Eingaben und der Ausgabe automatisiert extrahiert. Der große Vorteil dieser Methodik ist es, dass man das gelernte Modell $\hat{f}$ auch direkt auf neue Daten anwenden kann, also etwa auf bislang (dem Modell) unbekannte Tonaufnahmen. Der Nachteil ist jedoch, dass die manuelle Erzeugung von gelabelten Trainingsdatensätzen oft mit einem hohen Aufwand verbunden ist, der nicht automatisiert werden kann. Damit bleibt die Menge der verwertbaren Trainingsdaten natürlicherweise beschränkt.

Der Nachteil der Notwendigkeit von Labels für das Training fällt beim unüberwachten Lernen aus Abbildung 3 weg, da es hier keine Zielfunktion f gibt. Dennoch wird auch hier ein Trainingsprozess für das Modell $\hat{f}$ ausgeführt, welcher zum Ziel hat, die inneren Strukturen und eventuelle Korrelationen in den Daten möglichst gut zu beschreiben. Beispiele wären eine Warenkorbanalyse (Itemset Mining), also welche Objekte oft zusammen gekauft werden, Gruppenbildung (Clustering) in den Daten oder Graph-Mining Algorithmen. Auch im unüberwachten Lernen wird $\hat{f}$ meist iterativ bestimmt und die Güte der Approximation an die tatsächlich vorhandenen Strukturen in den Daten über einen Fehlerterm gemessen. Als Ergebnis vermittelt das Modell $\hat{f}$ explizites Wissen, welches bislang in den Daten nur immanent vorhanden war. Jedoch ist ein mittels unüberwachtem Lernen trainiertes Modell meist deskriptiver Natur und bringt damit den Nachteil mit sich, dass es nicht auf neue Daten angewendet werden kann.

Die Foundation-Modelle beruhen auf einer neueren Methodik, die als »selbstüberwachtes (self-supervised) Lernen« bezeichnet wird. Wie in Abbildung 4 dargestellt, wird ein Foundation-Modell $\hat{f}$ auf (an sich ungelabelten) Eingabedaten trainiert und kann dennoch auf neue Daten angewendet werden.

Namensgebend für die Methodik des selbstüberwachten Lernens ist, dass die Zielfunktion f die Labels aus den Daten gewissermaßen selbst erzeugt, so dass dann mit der bekannten Technik des überwachten Lernens aus Abbildung 3 ein Modell $\hat{f}$ trainiert werden kann, das die Zielfunktion f möglichst gut approximiert. Das einfachste Beispiel für eine solche Zielfunktion wäre die »Identität«, die besagt, dass die Eingabe möglichst korrekt reproduziert werden soll. Weitere Beispiele für einfache Zielfunktionen sind ausgelassene Wörter in einem Text, geschwärzte Ausschnitte in einem Bild oder den nächsten Frame in einer Videosequenz zu vorherzusagen. Auch weitere Transformationen der Eingaben sind möglich, die dann reproduziert werden müssen. Weitere komplexe Zielfunktionen f modellieren, ähnlich wie im unüberwachten Lernen, explizit Zusammenhänge und Korrelationen, die sich in den Daten befinden. Ein Beispiel ist das CLIP-Modell [RA21], welches die Zielfunktion lernt, ob ein Bild auf einer Webseite zusammen mit einer bestimmten Bildunterschrift zu finden ist oder nicht.

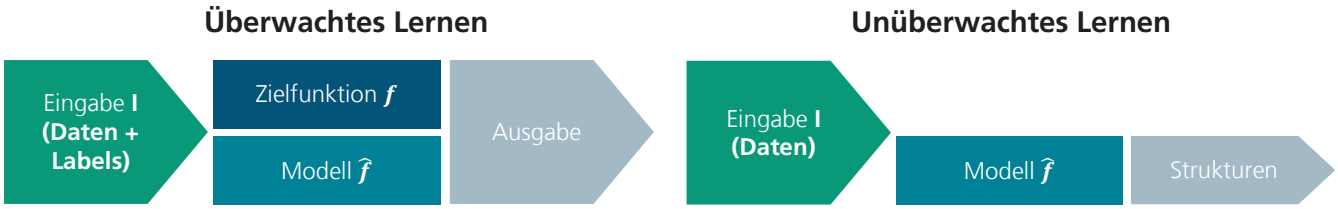


Abbildung 3: Unterscheidung zwischen überwachtem und unüberwachtem Lernen.

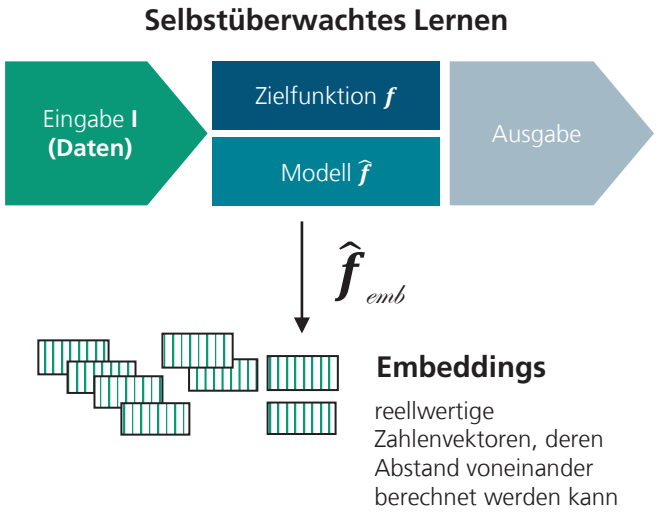


Abbildung 4: Selbstüberwachtes Lernen.

Ein wichtiger Bestandteil des Lernvorgangs für $\hat{f}$ besteht darin, dass eine Abbildung $\hat{f}_{emb}$ erstellt wird, welche die Eingabedaten auf sogenannte »Embeddings (Einbettungen)« abbildet, die lediglich reellwertige Zahlenvektoren sind (typischerweise mindestens der Länge 512). Eine Vielzahl von Untersuchungen und Experimenten hat jedoch gezeigt, dass Werte und Abstände der Embeddings eine semantische Bedeutung haben. Ein berühmtes frühes Beispiel konnte zeigen, dass für Wort-Embeddings $\hat{f}_{emb}$(»King«) - $\hat{f}_{emb}$(»Man«) = $\hat{f}_{emb}$(»Queen«) gilt [MSC13]. Gewissermaßen kann das Foundation-Modell $\hat{f}$ mit seiner zugehörigen Embedding-Funktion $\hat{f}_{emb}$ nahezu beliebige Eingabedaten im Raum der Embeddings semantisch korrekt verorten, sodass andere Anwendungen sie unmittelbar wiederverwenden können.

Als Beispiel für eine solche Anwendung sei hier eines der ersten generativen Bildmodelle DALL-E 2 [OA22] genannt, welches zu einer textuellen Beschreibung eines Bildes (z. B. »Astronaut auf einem Pferd«) passende Bilder erzeugen kann. Die Anwendung beruht auf dem bereits oben erwähnten Foundation-Modell CLIP. CLIP kann mit seiner Embedding-Funktion sowohl Bilder als auch Texte in demselben Embedding-Raum abbilden. Die herausfordernde Aufgabe für DALL-E 2 ist es nun, zu einem Text T ein in diesem Embedding-Raum ähnliches Bild B zu erzeugen. Hierfür nutzt DALL-E ein zweistufiges Verfahren. Zunächst wird basierend auf dem Text-Embedding $\hat{f}_{emb}$(T) ein plausibles Bild-Embedding $\hat{f}_{emb}$(B) in der Nähe von $\hat{f}_{emb}$(T) erzeugt (DALL-E Prior Modell). In einem zweiten Schritt erzeugt DALL-E in einem iterativen »Diffusionsprozess« ein Bild, das möglichst ähnlich zu dem vorgeschlagenen Bild-Embedding und damit auch dem Eingabetext ist. Auch eine Bilderzeugung direkt basierend auf $\hat{f}_{emb}$(T) wäre möglich, die Verwendung des Prior-Modells verbessert aber die Qualität und Variabilität der generierten Bilder.

Generative Sprachmodelle gehen noch einen Schritt weiter: mit der Embedding-Funktion wird zugleich eine Decodier-Funktion gelernt, die für gegebene Embeddings passende Texte erstellen kann. Damit wird das Sprachmodell in die Lage versetzt, zu einer gegebenen Eingabe die aus den Trainingsdaten heraus am »besten passende« Fortsetzung zu finden, wobei das »am besten passend« auf der einen Seite die Fortsetzung mit der höchsten Wahrscheinlichkeit unter allen möglichen Fortsetzungen ist. Auf der anderen Seite bezieht sich diese Fortsetzung aber aufgrund der reichhaltigen semantischen Bedeutung der Embeddings mehr auf den Sinn als auf die reine Wortfolge der Eingabe. Es konnte empirisch gezeigt werden [RA19], dass bei hinreichender Größe des Sprachmodells auch Eingabetexte wie »Übersetze von Deutsch nach Englisch« als Formulierung einer Aufgabe (Task) semantisch richtig abgebildet werden und somit die von der Bedeutung her wahrscheinlichste Fortsetzung eben die Übersetzung des nachfolgenden Eingabetextes auf Englisch ist. Somit sind also im Eingabetext sowohl die auszuführende Aufgabe als auch der Input für diese Aufgabe enthalten, ohne jedoch explizit voneinander in der Wortfolge getrennt sein zu müssen. Implizit hat das Foundation-Modell im Sinne eines Multi-Task Learnings viele verschiedene solcher Tasks durch den selbstüberwachten Trainingsprozess, die geschickt gewählte Zielfunktion und den dadurch semantisch sinnvoll gefüllten Einbettungsraum mitgelernt. Im überwachten Lernen hätte jede dieser Tasks als Teil einer Zielfunktion über Labels definiert werden müssen. Wie im oben besprochenen CLIP Modell lässt sich die grundsätzliche Fähigkeit, textuelle Aufgabenbeschreibungen semantisch richtig zu verorten, auch mit selbstüberwachten Lernverfahren im Bereich der Bildverarbeitung kombinieren. Ein Beispiel für automatisch mitgelernte »Downstream«-Aufgaben ist die Erzeugung von sogenannten synthetischen Trainingsdaten durch ein Foundation-Modell. So markiert das in [ME23] beschriebene Foundation-Modell auf einem Eingabebild aufgrund von Texteingaben wie »Personen, Fußgänger« die in dem Bild vorkommenden Fußgänger, bei einer Texteingabe wie »Autos, Räder« markiert es die Räder der im Bild vorkommenden Autos.

Zusammenfassend zeichnen sich Foundation-Modelle durch folgende Eigenschaften aus:

- Foundation-Modelle beruhen auf selbstüberwachtem Lernen und benötigen keine manuell erzeugten Labels.
- Foundation-Modelle können auf möglichst umfangreichen Datenmengen trainiert werden, weil keine Labels benötigt werden.
- Foundation-Modelle haben eine sehr große Anzahl an internen Parametern [AN23] – je nach Modell zwischen 100 Milliarden bis zu mehr als 500 Milliarden Parameter. Hierdurch können sie (Embedding-)Funktionen erzeugen, die komplexe semantische Eigenschaften der realen Daten in einfach handhabbaren, mathematischen Werten abbilden.
- Foundation-Modelle können in Eingaben formulierte Aufgaben verarbeiten und die Ergebnisse für diese Aufgabe für einen in der Eingabe mitgegebenen »eigentlichen

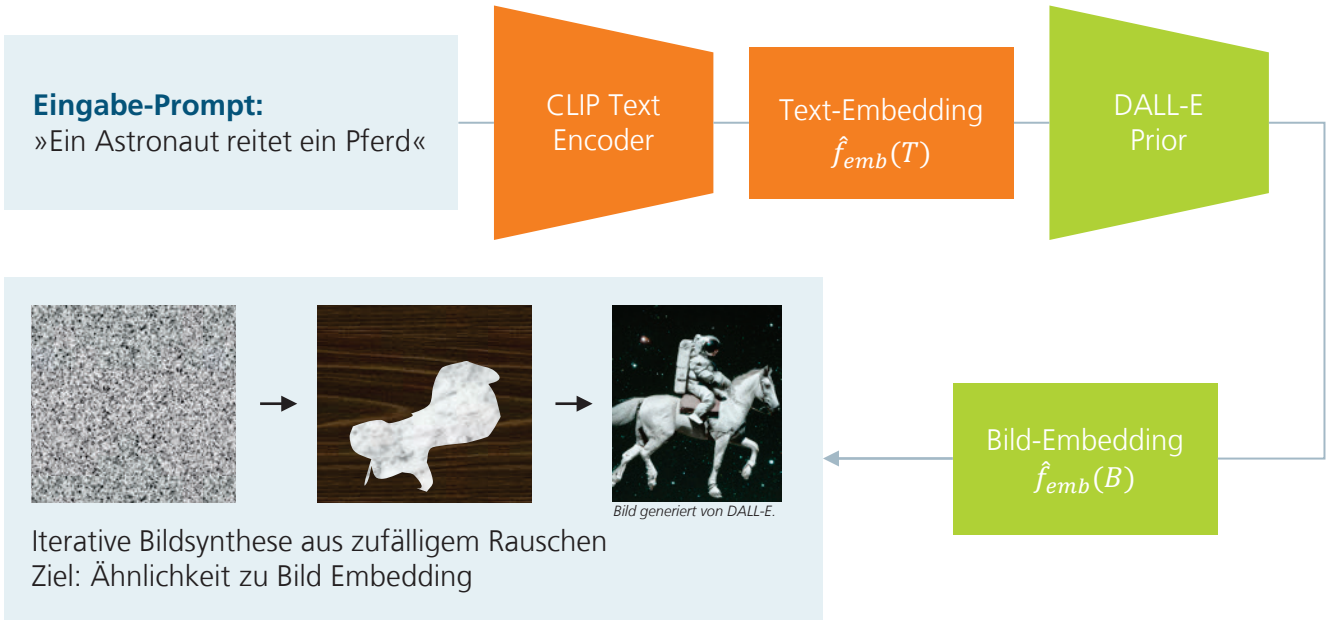


Abbildung 5: Erzeugung von semantisch zu einem Prompt passenden Bildern mit DALL-E 2.

Input« vorhersagen. Dabei wird in der Eingabe nicht explizit zwischen Aufgabe und eigentlichem Input unterschieden. Diese Fähigkeit haben sie implizit durch selbstüberwachtes Training gelernt [WE22a].

Letztendlich extrahieren Foundation-Modelle das immanent in den Daten vorhandene Wissen und erstellen Funktionen, mit denen direkt auf der semantischen Ebene mit den Daten gearbeitet werden kann. Dies erlaubt es, dass KI-Anwendungen auf neue Art und Weise entwickelt werden können, zum Beispiel auch ohne oder mit deutlich weniger gelabelten Daten. Die schnell wachsende Anzahl von Angeboten und Assistenzsystemen, die mit Hilfe von Foundation-Modellen bereits in kurzer Zeit auf den Markt gebracht wurden, verdeutlicht das enorme wirtschaftliche Potenzial dieser Technologie. Alleine in den ersten 6 Monaten nach der Einführung der Plugin-Technik für ChatGPT sind über 900 Plugins entwickelt und angeboten worden, die das Foundation-Modell bei Bedarf aufrufen kann, und es gibt »Prompt-Märkte«, in denen über 100.000 Prompts angeboten werden [WH23, ST23].

3 Vom Foundation-Modell zur KI-Anwendung

Durch das Training auf großen Datenmengen lernen Foundation-Modelle grundlegendes Wissen, das für viele Domänen und Aufgaben nützlich ist. Die Spezialisierung eines Foundation-Modells auf die konkrete Anwendungsdomäne und die eigentliche Aufgabe ist dabei eine wichtige Option, um ein Foundation-Modell für eine konkrete KI-Anwendung zu nutzen. Es werden anwendungsspezifische Daten und Domänenwissen gebraucht, um das in dem Foundation-Modell implizite Wissen über semantische Zusammenhänge für konkrete KI-Anwendungen nutzbar zu machen. Durch das bereits vorhandene Wissen des Foundation-Modells ist dieser Schritt deutlich vereinfacht im Vergleich zur Entwicklung einer KI-Anwendung ohne Foundation-Modelle. Es werden beispielsweise signifikant weniger oder überhaupt keine weiteren Daten benötigt. Die folgende Abbildung gibt einen Überblick über aktuell genutzte Möglichkeiten der Anwendungsentwicklung mit Foundation-Modellen.

Im Folgenden werden die verschiedenen Möglichkeiten erläutert, wobei die Reihenfolge sich grob an der Menge der benötigten zusätzlichen Daten orientiert. Dabei ist zu beachten, dass die dargestellten Möglichkeiten der Anwendungsentwicklung nicht nur Alternativen sind, sondern durchaus auch kombiniert werden können.

Programmiertes Prompt-Engineering

Als **Zero-Shot-Verfahren** werden Ansätze bezeichnet, die keine weiteren Daten zur Erstellung der Anwendung benötigen. Beispielsweise kann CLIP durch das erlernte Wissen über Bilder und Bildunterschriften ohne weitere Trainingsdaten genutzt werden, um Bilder in vorgegebene Klassen einzuteilen. Dabei wird jede Klasse über einen Eingabetext (einen Prompt)

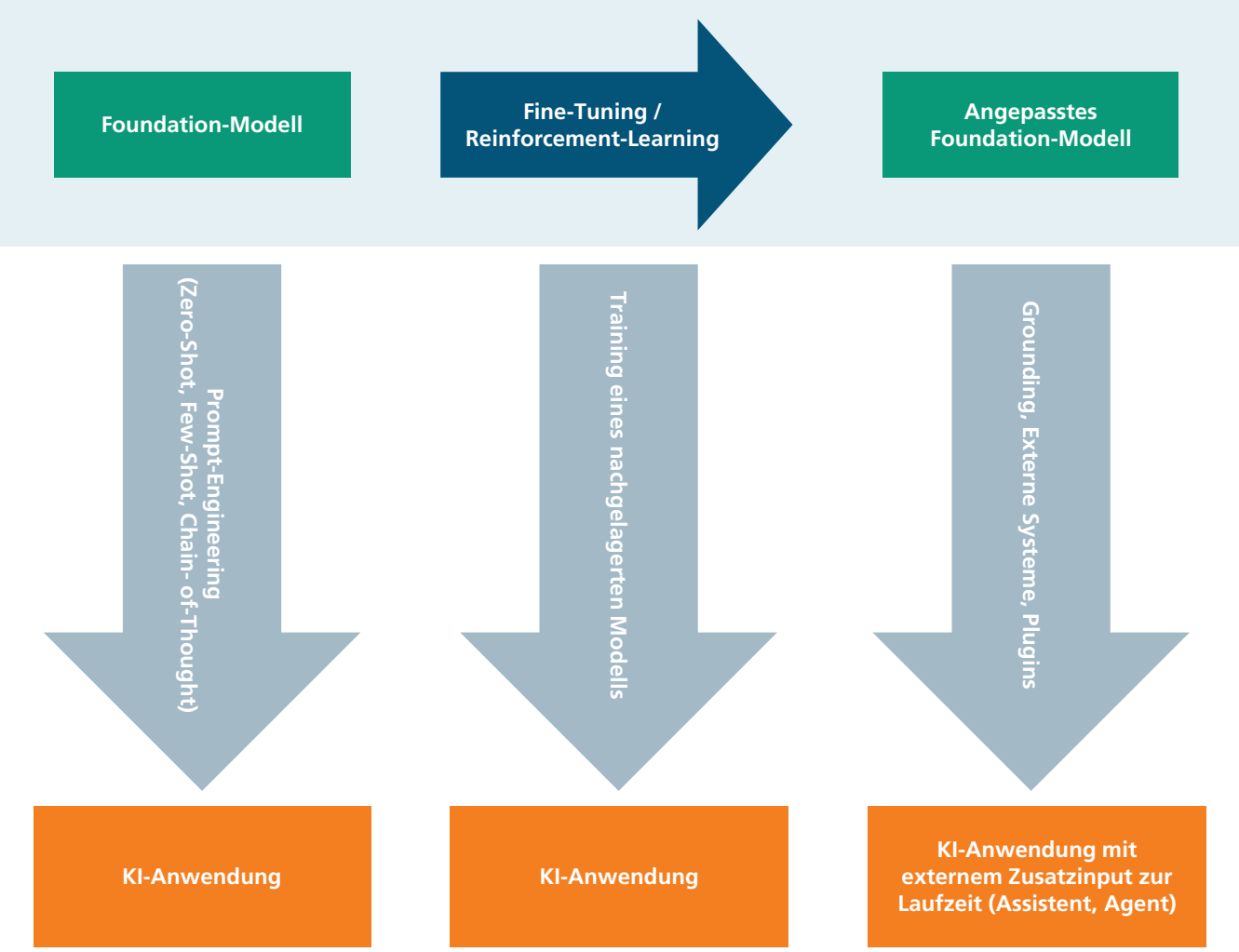


Abbildung 6: Übersicht über Verfahren zum Einsatz von Foundation-Modellen zur Entwicklung von KI-Anwendungen.

beschrieben, zum Beispiel »Ein Bild von einem Hund«, »Ein Bild von einer Katze«, »Ein Bild von einem Haus«. Dann wird über die Embedding-Funktion bestimmt, für welchen dieser »Prompts« der Abstand $\boldsymbol{f}_{\text{emb}}$(prompt) zu $\boldsymbol{f}_{\text{emb}}$(bild) minimal wird. Die zugehörige Klasse wird dann als Ergebnis der Anwendung ausgegeben. Auch in reinen Sprachmodellen können Zero-Shot-Verfahren zur Anwendung kommen. Hierbei wird der in Kapitel 2 erläuterte Zusammenhang ausgenutzt, dass sowohl die in einer Eingabe enthaltene Aufgabe als auch der eigentliche Input für die Aufgabe semantisch im Embedding-Raum repräsentiert werden. Beispielsweise enthält die Eingabe »Wie alt wurde Albert Einstein?« die Aufgabe, das Alter einer Person zu bestimmen, als auch die Angabe, die Person zu bestimmen. Dementsprechend kann das Modell die Antwort als die »höchstwahrscheinliche Fortsetzung« vorhersagen.
Bei den Zero-Shot-Verfahren wird die eigentliche Aufgabe (oder gesuchte Klasse im Beispiel von CLIP) textuell über einen Prompt formuliert, welcher dann in den Embedding-Raum abgebildet wird. Es hat sich gezeigt, dass die derzeitigen Foundation-Modelle sehr sensitiv auf die konkrete Formulierung dieses Prompts reagieren. Nach heutigem Stand gehört viel Ausprobieren und dadurch gesammelte Erfahrung mit einem Foundation-Modell dazu, um für eine gegebene Anwendung gut funktionierende Prompts zu entwickeln, die die Aufgabe für das Foundation-Modell genau genug definieren. Dieses Verfahren wird oft als **Prompt-Engineering** bezeichnet, da es keine klassische Entwicklungsarbeit, sondern mehr eine Aktivität seitens der Benutzer*innen darstellt. Um zu prüfen, welche Prompts »gut« oder auch »gut genug« sind, wird in den Zero-Shot-Verfahren die eigentliche KI-Anwendung auf Testdatensätzen getestet. Somit werden auch bei den Zero-Shot-Verfahren weitere, meist gelabelte Datensätze benötigt – wenn auch nur, um die Anwendung zu testen. Neben dem Prompt-Engineering zur Entwicklung von KI-Anwendungen wird dieses Verfahren natürlich in der interaktiven Nutzung einer generativen KI eingesetzt. In diesem Fall stellt sich jedoch weniger die Frage nach der allgemeinen Gültigkeit der gefundenen Prompts für nachfolgende Aufgaben.

Eine Fortentwicklung des Zero-Shot-Verfahrens wird als **Few-Shot-Learning** bezeichnet. Hierzu wird im Sinne des Prompt-Engineerings die zu lösende Aufgabe – neben der Aufgabenbeschreibung und dem eigentlichen Input – noch durch die Angabe von weiteren, gelabelten Datensätzen spezifiziert. Das bedeutet, dass eine fest ausgewählte, kleine Menge (z. B. 10 korrekte Frage-Antwort-Paare) an gelabelten Daten aus dem Bereich der zu entwickelnden Anwendung als Kontext in die Eingabe gegeben wird. Es konnte empirisch auf großen Testdatensätzen gezeigt werden, dass damit die Treffergenauigkeit der Ausgaben von Foundation-Modellen deutlich erhöht werden kann. Dieses Verfahren ist besonders effizient, wenn die Eingaben und möglichen Ausgaben stark standardisiert sind, beispielsweise beim Beantworten von Multiple-Choice-Fragen.

Eigentlich ist die Bezeichnung »Learning« im »Few-Shot-Learning« irreführend, da das bestehende Foundation-Modell nicht modifiziert wird, also kein weiterer Trainingsprozess stattfindet. Es wird lediglich die von dem Foundation-Modell zu erfüllende Aufgabe (siehe Kapitel 2) durch die Angabe von Beispielen in der Eingabe genauer beschrieben. Ein Beispiel für ein »One-Shot-Learning« wäre die Eingabe

»Translate from English to French:
dog =>chien, cheese=>«

in der ein gelabeltes Beispiel für die zu erfüllende Aufgabe gegeben wird. Die erwartete Ausgabe ist folglich die Übersetzung von »cheese« ins Französische, also »fromage«.
Insgesamt ist das Few-Shot-Verfahren ein wichtiger Baustein für eine Reihe von Techniken, mithilfe derer die Eingabe des Foundation-Modells mit einer geringen Menge an Daten um zusätzliche Kontextinformationen angereichert werden kann.

Das sogenannte **Chain-of-Thought** (Gedankengang-)Verfahren [WE22] nutzt die Technik des Few-Shot-Verfahrens, um die zu lösende Aufgabe nicht nur durch die Angabe von korrekten Frage-Antwort-Paaren zu beschreiben, sondern auch die logischen Schritte und Schlussfolgerungen zur Erzeugung der korrekten Antwort als Teil der Eingabe explizit darzulegen. Die Idee des Chain-of-Thought-Verfahrens ist somit, das Modell durch Angabe von Beispielen in der Eingabe anzuleiten, neben der Antwort auch noch die zugrundeliegenden Schlüsse explizit darzulegen. Auch ist es möglich, das Modell aufzufordern, die Gedankenschritte in der Antwort anzugeben (z. B. durch die Formulierung »denke Schritt für Schritt«), ohne dass die Teilschritte als Eingabe formuliert werden müssen. Es konnte gezeigt werden, dass damit die Genauigkeit der Ausgabe von Foundation-Modellen, zum Beispiel bei der Lösung von arithmetischen Textaufgaben auf entsprechenden Testdatensätzen, deutlich erhöht werden kann [SU22]. Im Prinzip wird mit diesem Verfahren auch die Möglichkeit eröffnet, Wissen über die Methodik des Problemlösens in einer Anwendungsdomäne in die Anwendungsentwicklung mit einzubringen. Jedoch sind die Erfahrungswerte über die Generalisierungsfähigkeit in realen Anwendungsdomänen noch beschränkt.

Training eines nachgelagerten Modells

Die bislang beschriebenen Verfahren der Anwendungsentwicklung verwenden keine oder nur wenige gelabelte Daten und trainieren keine neuen Modelle oder angepasste Foundation-Modelle. Liegen mehr gelabelte Daten vor, so ergeben sich weitere Möglichkeiten. Ein einfaches Verfahren ist das **Training eines nachgelagerten Modells**. Das Grundprinzip lässt sich einfach am Beispiel der Entwicklung einer Anwendung zur Bildklassifizierung unter Nutzung von CLIP erläutern. Wie im obigen Abschnitt zu Zero-Shot-Verfahren beschrieben, ist

es Ziel der Anwendung, ein gegebenes Bild einer von mehreren gegebenen Klassen zuzuordnen (z. B. »Hund«, »Katze«, »Haus«). Es liegen gelabelte Trainingsdaten für das Trainieren eines überwachten Modells vor. Anstatt dieses Modell $\hat{f}$ nun direkt auf den Bilddaten zu trainieren, wird zunächst die Embedding-Funktion von CLIP auf die Bilder angewendet, es werden also die Paare (f_{emb}(bild), label) als Trainingsdaten genutzt. Dementsprechend erhält man die Vorhersage für ein neues Bild nun als $\hat{f}(f_{emb}(\text{bild}))$. Es konnte auf großen Testdatensätzen gezeigt werden, dass durch dieses Verfahren sowohl die benötigte Komplexität von $\hat{f}$ als auch die für das Training benötigte Datenmenge gegenüber bestehenden, von Grund auf überwacht trainierten tiefen neuronalen Netzen zur Bildklassifikation um Größenordnungen reduziert werden konnte, ohne Einbußen an der Treffergenauigkeit der Vorhersagen in Kauf nehmen zu müssen.

Fine-Tuning und Reinforcement-Learning durch Feedback

Das beschriebene Verfahren zum Training eines nachgelagerten Modells erfordert keine Veränderung des genutzten Foundation-Modells und ist somit relativ einfach durchzuführen. Aufwändiger in Bezug auf die genutzte Infrastruktur ist in der Regel das sogenannte **Fine-Tuning**. Die grundsätzliche Technik des Fine-Tuning ist schon lange etabliert, zum Beispiel im Bereich der Bildverarbeitung, wo vortrainierte tiefe Netze als Backbone eingesetzt werden, die zum Beispiel auf allgemeinen Bilddaten zur Objekterkennung trainiert worden sind. Im Fine-Tuning werden sie dann auf speziellen (gelabelten) Trainingsdaten der Anwendungsdomäne weiter trainiert. Genau so kann ein Foundation-Modell basierend auf anwendungsspezifischen Daten weiter trainiert und die internen Parameter des Modells diesbezüglich optimiert werden. Diese Technik wird auch von den Herstellern von Foundation-Modellen zusätzlich zum selbstüberwachten Lernen genutzt, um ein Sprachmodell daraufhin zu trainieren, Anweisungen im Prompt besser zu verstehen und zu verfolgen. Des Weiteren werden auch durch gezielt vorab ausgeführte Prompts die Ausgaben für nachfolgende Abfragen beeinflusst und gesteuert (sogenanntes Instruction Fine-Tuning). »Reinforcement-Learning by Human Feedback« nennt sich eine weitere Methode, mit der Hersteller ihrem Modell anzutrainieren versuchen, von Menschen bevorzugte Antworten zu produzieren, also etwa höflich und hilfsbereit zu sein und keine unerlaubten Informationen herauszugeben [GL22]. Man spricht in diesem Zusammenhang von »Alignment«. Bei dieser Methode wird aus menschlichem Feedback ein separates Bewertungsmodell trainiert und das Foundation-Modell gelegentlich so trainiert, dass es besser bewertete Antworten produziert.

Grounding und Plugins

Die bislang beschriebenen Techniken der Anwendungsentwicklung werden in der Design- und Entwicklungsphase der eigentlichen KI-Anwendung eingesetzt. Dementsprechend ergibt sich daraus eine auf einen Anwendungszweck und einen Anwendungsbereich spezialisierte KI-Anwendung, die prinzipiell vor ihrer Inbetriebnahme auf ihre Eigenschaften, insbesondere in Bezug auf die Dimensionen der Vertrauenswürdigkeit, getestet werden sollte. Darüber hinaus können jedoch auch Anwendungen entwickelt werden, die die Technik des Prompt-Engineerings nutzen, um zur Laufzeit der KI-Anwendung weitere, anwendungs- und situationsabhängige Informationen von extern zu beschaffen, die letztendlich die von dem Foundation-Modell zu lösende Aufgabe im konkret vorliegenden Fall der aktuellen Anfrage möglichst präzisieren. Beim sogenannten **Grounding** [MC23, NA21] ist die KI-Anwendung so programmiert, dass sie zur Laufzeit Abfragen an externe Systeme durchführt, deren Ergebnisse dann nach einer anwendungsspezifischen Algorithmik in das Prompt mit eingepflegt werden, ehe der Aufruf des Foundation-Modells erfolgt. Dabei sind den Möglichkeiten, welche externen Systeme genutzt werden, prinzipiell keine Grenzen gesetzt. Dies kann von einer einfachen Suchanfrage bei einer Internet-Suchmaschine über anwendungsabhängige hochqualitativ gepflegte Datenbanksysteme bis zur spezialisierten Dokument- und Wissensdatenbank gehen. Es sind sogar Aufrufe anderer Foundation-Modelle denkbar.

Neben der Nutzung externer Systeme können Anwendungen Informationen auch aus der langfristigen Interaktion mit den Benutzer*innen gewinnen und die Eingabe an das Foundation-Modell damit anreichern.

Eine noch höhere Dynamik zur Laufzeit kann über sogenannte **Plugins** erreicht werden. Ein Plugin kann, wie im Grounding, über externe Systeme Information abrufen. Dabei generiert das Foundation-Modell selbst den Plugin-Aufruf, den die umgebende Software erkennt, ausführt und durch das Ergebnis im Eingabetext ersetzt. Neben der reinen Informationsbeschaffung können auch Aktionen (wie z. B. das Senden einer E-Mail) über Plugins aufgerufen werden. Auf diese Art können oft als **»autonome Agenten«** bezeichnete Anwendungen entstehen. Diese autonomen Agenten gestalten die Interaktion mit der Umwelt dynamisch aus den Laufzeit-Ausgaben des zugrundeliegenden Foundation-Modells heraus.

4 Risiken von KI-Anwendungen und Foundation-Modellen

Verschiedene Ansätze und Frameworks verfolgen das Ziel, die Vertrauenswürdigkeit von KI-Anwendungen, auch im Hinblick auf die bevorstehende europäische KI-Verordnung, bewertbar zu machen und Richtlinien zu geben, die bei der Entwicklung und dem Betrieb von vertrauenswürdigen KI-Anwendungen zu beachten sind [PO21, LE21, VD22, OE22]. Dabei werden typischerweise immanente KI-Risiken, zum Beispiel in Bezug auf die Zuverlässigkeit oder Sicherheit der Anwendung oder in Bezug auf möglicherweise diskriminierende Auswirkungen, systematisch in verschiedenen Dimensionen der Vertrauenswürdigkeit untersucht.

Auch wenn der Hersteller eines Foundation-Modells durch Fine-Tuning bereits umfangreiche Absicherungsmaßnahmen ergriffen hat (siehe Erläuterung in Kapitel 3), macht die Nutzung von Foundation-Modellen in KI-Anwendungen eine systematische Untersuchung bezüglich anwendungsspezifischer KI-Risiken natürlich nicht obsolet. Jedoch ergeben sich möglicherweise weitere Herausforderungen [BS23, LE23, WA23], die sich wiederum auf die KI-Anwendung auswirken. Abbildung 7 fasst spezielle Risiken bei der Nutzung von Foundation-Modellen entlang der in [CR19, PO20, PO21] beschriebenen Dimensionen der Vertrauenswürdigkeit zusammen und bildet sie auf die Bedeutung in einer KI-Anwendung ab. Hierbei wird deutlich, dass es unmöglich ist, ein Foundation-Modell für alle denkbaren Anwendungskontexte abzusichern und vertrauenswürdig zu gestalten, sondern dass Risiken immer im Anwendungskontext betrachtet werden müssen. Die folgende Darstellung der Dimensionen der Vertrauenswürdigkeit orientiert sich an dem von Fraunhofer IAIS entwickelten »KI-Prüfkatalog zur Gestaltung vertrauenswürdiger Künstlicher Intelligenz« [PO21] und beschreibt für die einzelnen Risikodimensionen Implikationen durch die Nutzung von Foundation-Modellen in KI-Anwendungen.

Fairness

Die Dimension Fairness soll sicherstellen, dass die KI-Anwendung nicht zu ungerechtfertigter Diskriminierung und Stereotypisierung 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. Da die Foundation-Modelle nicht auf vollständig kuratierten, sondern auf möglichst vielen (allen) Daten trainiert werden, werden Ungleichverteilungen in den bestehenden Daten zunächst teilweise übernommen. Es ist bekannt, dass hier insbesondere Problematiken bzgl. Bevorzugung/Benachteiligung von

Dimension der Vertrauenswürdigkeit	Bedeutung in der KI-Anwendung	Spezielle Risiken bei der Nutzung von Foundation-Modellen
Fairness	Behandelt die KI alle Betroffenen fair?	- Stereotypisierung, Diskriminierung
Autonomie & Kontrolle	Ist eine selbstbestimmte, effektive Nutzung der KI möglich?	- Selbstsichere Ausdrucksweise verleitet zu übermäßigem Vertrauen (blind trust) - Verkümmern von selten genutzten Fähigkeiten (enfeeblement) - Autonome Agenten, die sich unerwartet entwickeln - Emotionale Abhängigkeiten
Transparenz	Sind Funktionsweise und Entscheidungen der KI nachvollziehbar?	- Künstlich generierte Inhalte sind schwer zu erkennen - Erfundene Begründungen
Verlässlichkeit	Funktioniert die KI verlässlich?	- Fehlende Aktualität - Toxische und andere verbotene Inhalte - Halluzinationen, Fehlinformationen, Generierung von Code mit Fehlern - Zugriff auf externe Systeme (plugins)
Sicherheit	Ist die KI sicher gegenüber Angriffen, Unfällen und Fehlern?	- Desinformationen, Deepfakes, Generierung von Schadcodes - Personalisierter Betrug - Angriffe durch unerwartete oder manipulative Eingaben
Datenschutz	Schützt die KI-Anwendung sensible Informationen?	- Ein- oder Ausgabe sensitiver Inhalte - Nutzung rechtlich geschützter Inhalte

Abbildung 7: Spezielle Risiken von Foundation-Modellen in den Dimensionen der Vertrauenswürdigkeit.

geschlechtsspezifischen oder ethnischen Gruppen eine Rolle spielen und zu **Diskriminierung und Stereotypisierung** führen können. Beispiele sind antimuslimische Vervollständigungen in GPT-3 [AB21] oder die Bildgenerierung von DALL-E, die, bevor Gegenmaßnahmen ergriffen wurden, beispielsweise unter geschlechtsneutralen Eingaben wie »heroic firefighter« oder »CEO« vorwiegend männliche Ausgaben produzierte [OB22]. Auch wenn die Hersteller von Foundation-Modellen solche Fairness-Defizite zum einen explizit bekannt geben und diesen zum anderen zum Beispiel mit Filtertechniken entgegenwirken, bleibt festzuhalten, dass die Nutzung eines Foundation-Modells die Fairnessproblematik eher verstärken als abmindern kann. Diesbezügliche Tests auf Anwendungsebene bleiben unabdingbar.

Autonomie und Kontrolle

Dass eine KI-Anwendung die Autonomie des Menschen nicht beinträchtigen soll, ist bereits in den ethischen Grundprinzipien der HLEG-Kommission verankert [EU19], an denen sich auch die geplante KI-Verordnung orientiert. Der »KI-Prüfkatalog« des Fraunhofer IAIS operationalisiert diese grundsätzliche Anforderung auf der Ebene des Designs und des Zwecks der KI-Anwendung. Zunächst ist hier zu beurteilen, welcher **Grad an Autonomie** für den Zweck der KI-Anwendung angemessen ist. Anschließend wird untersucht, ob der Mensch durch die KI-Anwendung angemessen unterstützt wird und ausreichend Handlungsspielraum in der Interaktion mit der KI-Anwendung erhält. Auch in Bezug auf Foundation-Modelle ist diese Bewertung natürlich nur im Kontext einer konkreten abgeleiteten KI-Anwendung möglich, da hier erst der Zweck und die Möglichkeiten der Interaktion mit dem Menschen festgelegt werden. In diesem Sinne bleibt diese Dimension auch beim Einsatz von Foundation-Modellen weiterhin prüfungsrelevant. Dabei ist auch zu berücksichtigen, dass die bekannten Risiken bezüglich **»blindem Vertrauen«** (blind trust) und **»emotionaler Abhängigkeit«** gerade bei einer menschenähnlichen, oftmals selbstsicher wirkenden Ausdrucksweise von Sprachmodellen verstärkt auftreten können. Zudem erhöht sich durch die Nutzung leistungsfähigerer Foundation-Modelle das Risiko von **»Enfeeblement«** (die Verkümmerung selten genutzter Fähigkeiten), da der Mensch Gefahr läuft, viele Aufgaben an KI-Modelle abzugeben, und dementsprechend selbst Fähigkeiten verlernt (z. B. das Übersetzten von Texten, was sich durch Sprachmodelle sehr schnell und einfach realisieren lässt) [PA23]. Auf einer gesamtgesellschaftlichen Ebene können sich weitere Herausforderungen ergeben, wie die zu erwartende Umwälzung der Arbeitswelt oder die gezielte Produktion und Verbreitung von Falschinformation. Diese Herausforderungen sind auch durchaus konkreter, als das in [SA23] erwähnte Risiko der »Ausrottung der Menschheit«, welches alle namhaften Hersteller explizit nicht eingehen wollen.

Transparenz

Unter dem Oberbegriff Transparenz sind Aspekte der Dokumentation und Information, Nachvollziehbarkeit sowie Erklärbarkeit subsumiert. Die Dimension Transparenz untersucht insbesondere, ob die grundlegende Funktionsweise der KI-Anwendung sowohl für Anwender*innen als auch für Entwickler*innen angemessen nachvollziehbar ist, und ob Ergebnisse der KI-Anwendung reproduziert, möglicherweise begründet und gegebenenfalls im Sinne einer Auditfähigkeit belegt werden können.

Die Transparenzdimension wird in Foundation-Modellen meist bereits anbieterseitig auf der Ebene der Dokumentation und Beschreibung der Daten und Modelle adressiert. Zum Beispiel dokumentieren Modelcards [MI19] die Ausgaben eines Modells in unterschiedlichen Szenarien. In Bezug auf ihren tatsächlichen Einsatz in konkreten Anwendungen bedarf es zusätzlich einer systematischen Herangehensweise. Gerade die Tatsache, dass **künstlich generierte Inhalte** oftmals schwer als solche zu erkennen sind, macht es umso wichtiger, dass die Nutzer*innen hinreichend über den Einsatz einer KI informiert sind. Dies ist auch eine der Grundforderungen der geplanten EU KI-Verordnung (siehe Kapitel 5). In Bezug auf die Nachvollziehbarkeit verstärken Foundation-Modelle zum einen aufgrund ihrer schieren Größe die bereits von herkömmlichen tiefen neuronalen Netzen bekannten Probleme, bieten aber durch ihre reichhaltige interne semantische Struktur im Embedding-Raum auch völlig neue Möglichkeiten. Im simpelsten Fall kann man aufgrund der in Kapitel 2 dargestellten impliziten Fähigkeit der Foundation-Modelle zum »Multi-Task Learning« neben der eigentlichen Ausgabe auch noch eine Begründung oder Erklärung (z. B. Quellenangaben) für die Ausgabe mit ausgeben lassen. Auch Promptingtechniken wie Chain-of-Thought (siehe Kapitel 3) erzeugen Ausgaben, die die Nachvollziehbarkeit für die Anwender*innen erhöhen. Natürlich unterliegen alle diese zusätzlichen erklärenden Ausgaben den gleichen prinzipiellen Vorbehalten in Bezug auf Korrektheit und Vollständigkeit, wie die eigentliche Modellausgabe (sie könnten also auch **»erfundene Begründungen«** sein). Es gehört zum transparenten Design der eigentlichen KI-Anwendung, Nutzer*innen auch darüber angemessen zu informieren. Neben dieser naheliegenden einfachen Erklärungsmöglichkeit auf Nutzerebene bieten Foundation-Modelle aber auch das Potenzial, im Embedding-Raum systematisch und algorithmisch nach Zusammenhängen und Erklärungen zu suchen.

Verlässlichkeit

Die Dimension Verlässlichkeit umfasst die Qualität der KI-Komponente in Bezug auf verschiedene Aspekte: Performanz, Robustheit, d. h. die Konsistenz ihrer Ausgaben unter kleinen Veränderungen der Eingabedaten, die Einschätzung der Modellunsicherheiten sowie das Abfangen von Fehlern.

Die Nutzung von Foundation-Modellen hat maßgeblichen Einfluss auf die verschiedenen Aspekte der Verlässlichkeit einer KI-Anwendung in einem bestimmten Anwendungskontext. Während Foundation-Modelle an vielen Stellen zu performanteren und robusteren Ergebnissen führen, die bereits durch die Anbieter*innen getestet werden, treten unentdeckte Verzerrungen und »spurious correlations«[4] teilweise erst im Anwendungskontext auf [BO21]. Foundation-Modelle neigen dazu, immer wieder faktisch inkorrekte, aber plausibel klingende Antworten zu produzieren. Diese Neigung zu sogenannten **Halluzinationen** wird dann besonders problematisch, wenn die Ausgaben nur mit Expert*innenwissen als falsch bewertet werden können. Die Erfindung nicht existenter juristischer Zitationen [RE23] ist beispielsweise besonders in Kombination mit der selbstsicheren Ausdrucksweise problematisch. Zusätzlich sind Ausgaben teils inkonsistent und variieren, zum Beispiel abhängig von der Wortreihenfolge der Eingabe. Falsche Ausgaben entstehen mitunter auch durch **fehlende Aktualität** der Foundation-Modelle: Da die Trainingsdaten nur einen in der Vergangenheit liegenden Zeitraum umfassen, reicht auch der Wissensstand eines Modells bis zu diesem bestimmten Zeitpunkt. Eine weitere Herausforderung stellt die unerwünschte Generierung von **toxischen und verbotenen Inhalten** dar, welche von Foundation-Modellen aufgrund des immensen und teilweise ungefilterten Wissensschatzes reproduziert werden.

Um diese Herausforderungen zu adressieren, wird in verschiedenen Ansätzen (z. B. »Safety Fine-Tuning« [UN22]) untersucht, wie unerwünschte oder falsche Ausgaben möglichst vermieden werden können. Beispielsweise werden bereits durch Anbieter Filter eingebaut, Grounding eingesetzt oder aktuelle Informationen hinzugezogen. Durch den **Zugriff auf externe Systeme** zur Laufzeit kann die Verlässlichkeit und Aktualität zwar erhöht werden, führt aber eine weitere Abhängigkeit von eben diesen Systemen ein. Auch Nutzer*innen können durch die Angabe von Kontext und Beispielen (Few-Shot-Learning), die Untersuchung mehrerer Lösungspfade (»Self-Consistency« [WA22]) oder das Abfragen von Zwischenschritten (Chain-of-Thought-Prompting [WE22]) die Verlässlichkeit der Ausgaben beeinflussen.

Sicherheit

Die Dimension Sicherheit adressiert sowohl die Absicherung der KI-Anwendung gegenüber Angriffen und Manipulationen als auch Eigenschaften der funktionalen Sicherheit. Da sich die Maßnahmen in dieser Dimension primär auf die Einbettung der KI-Komponente beziehen, können für Foundation-Modelle viele der bisherigen Sicherheitsmechanismen für KI-Anwendungen (wie z. B. klassische Methoden der IT-Sicherheit oder das Einbauen eines Fail-Safe-Modus) übernommen werden. Dennoch verstärken oder verändern sich durch die Verwendung von Foundation-Modellen teilweise die Sicherheitsrisiken. Durch die Nutzung von nicht kuratierten öffentlichen Daten aus dem Web ohne direkte Trainingsüberwachung werden beispielsweise neue Möglichkeiten für **Data-Poisoning**-Attacken auf Foundation-Modelle geschaffen. Weiterhin ergeben sich Risiken dadurch, dass Foundation-Modelle oftmals im Betrieb weiterlernen und sensitiv auf Eingaben des Nutzers reagieren. **Prompt Injection Attacks** zielen durch entsprechende **manipulative Eingaben** auf die gezielte Produktion falscher oder verletzender Inhalte ab. Zum Beispiel lassen sich Schutzfilter in gängigen Foundation-Modellen umgehen, indem man das Foundation-Modell dazu auffordert, aus der Perspektive einer bestimmten Persönlichkeit zu antworten. Ein weiteres Risiko ist der Missbrauch von Foundation-Modellen für betrügerische oder kriminelle Zwecke durch **Deepfakes** zur Rufschädigung oder zum personalisierten Betrug und der Generierung von Desinformationen oder Schadcodes. Das Risiko von fehlender Nutzerinformation und Täuschung kann ebenfalls als Teil der Dimension Autonomie und Kontrolle betrachtet werden.

Das Risikogebiet der funktionalen Sicherheit adressiert Risiken, aus denen eine Gefährdung der Außenwelt aufgrund von Fehlern oder Unfällen der KI-Anwendung folgt. Durch die Nutzung von mächtigen Foundation-Modellen werden KI-Anwendungen für komplexere und verantwortungsvollere Aufgaben eingesetzt, woraus oftmals ein stärkerer Schutzbedarf hinsichtlich der funktionalen Sicherheit folgt.

Datenschutz

Die Dimension Datenschutz 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 oder lizenzgebundener und urheberrechtlich geschützter Daten adressiert. Foundation-Modelle nutzen große Mengen von öffentlichen sowie geschützten Daten. Außerdem lernen viele Foundation-Modelle auf in den Eingaben bereitgestellten Informationen weiter. Auch in der Dimension Datenschutz verstärken sich so bestehende Risiken, wie die Extraktion von Trainingsdaten oder die Verknüpfung von Daten. Es sind verschiedenen Fälle bekannt, bei denen das Modell sensitive Inhalte aus Eingabedaten lernt und zufällig an anderer Stelle wieder ausgibt. Durch **Model Inversion Attacken** können beispielsweise anhand

4 Zufällige statistische Zusammenhänge, die aber inhaltlich falsch sind, werden als »spurious correlations« bezeichnet. Ein bekanntes Beispiel ist, dass in Regionen mit mehr Störchen auch mehr Kinder geboren werden. Da aber das Training eines KI-Modells nur auf statistischen Zusammenhängen beruht, können solche »falschen« Zusammenhänge zu falschen Ausgaben in der KI-Anwendung führen.

gezielter und systematischer Abfragen des Modells **sensible Daten** wie Sozialversicherungsnummern oder realistische Abbildungen (vorher unbekannter) Personen als Ausgabe erzeugt werden [CA21]. Ebenfalls kann so beispielsweise die Struktur der Trainingsdaten in vielen Fällen zurückgewonnen werden bzw. Daten, die den Trainingsdaten ähneln, künstlich erzeugt werden. Ein Ursprung dafür ist das Overfitting dieser Foundation-Modelle an oftmals heterogene Trainingsdaten [FE20]. Ein weiteres Risiko im Bereich Datenschutz ist die Nutzung rechtlich geschützter Inhalte in Trainingsdaten von Foundation-Modellen. Dies führt potenziell dazu, dass Foundation-Modelle diese Trainingsdaten als Teil der Modellausgabe reproduzieren und damit Urheber- oder Eigentumsrechte verletzen. Beispielsweise werden basierend auf Trainingsdaten Inhalte generiert, die dem Stil gewisser Künstler*innen nachempfunden sind oder Textstellen aus einem Buch wiedergeben [HE23].

Während durch die Nutzung von Foundation-Modellen einerseits Datenschutzrisiken verstärkt werden, ermöglicht die Nutzung von vortrainierten Foundation-Modellen in KI-Anwendungen andererseits oftmals auch eine Reduzierung der Menge an benötigten vertraulichen Daten. Weiterhin zeigt die Erforschung und Bereitstellung von Ansätzen wie »private GPT« datenschutzfreundlichere Alternativen zu bekannten großen Foundation-Modellen auf.

Auch wenn sich die Risiken systematisch verschiedenen Risikodimensionen zuordnen lassen, sind diese oftmals nicht unabhängig voneinander und unterliegen teilweise sogar Trade-offs. Die Gesamtbewertung der Vertrauenswürdigkeit einer KI-Anwendung erfordert somit eine abwägende Gesamtbetrachtung über die verschiedenen Risikodimensionen hinweg.

5 Europäische KI-Verordnung und Foundation-Modelle

Aufgrund der Risiken von KI-Anwendungen und ihrer schnellen Verbreitung rückte in den vergangenen Jahren zunehmend die Notwendigkeit von Regulierung in den Fokus. Die weltweit erste, umfassende Regulierung von Künstlicher Intelligenz ist die europäische KI-Verordnung (AI Act), die zum Zeitpunkt der Veröffentlichung dieses Whitepapers kurz vor der formellen Verabschiedung steht. Im Folgenden werden sowohl die Anforderungen an KI-Anwendungen, als auch die Anforderungen an Foundation-Modelle betrachtet, welche voraussichtlich 2 Jahre bzw. 12 Monate nach dem Inkrafttreten der KI-Verordnung Anwendung finden werden [KOM23]. Da die Trilogverhandlungen zwischen EU Parlament, Rat und Kommission eine vorläufige Einigung erzielt haben, deren fachliche Einzelheiten zum Zeitpunkt der Veröffentlichung dieses Whitepapers ausgearbeitet werden [RAT23], beziehen sich die Ausführungen in diesem Kapitel auf dem Kompromissvorschlag des EU Parlaments von Juni 2023 [EU23]. Dieser enthält als einziger Entwurf zur europäischen KI-Verordnung detaillierte Regelungen zu Foundation-Modellen.

Die KI-Verordnung verfolgt grundsätzlich einen risikobasierten Ansatz, der KI-Anwendungen je nach Anwendungszweck in vier Risikoklassen einteilt: minimal, mäßig, hoch und unvertretbar. Anwendungen mit minimalem Risiko, wie beispielsweise Spamfilter, werden nicht reguliert. Anwendungen mit mäßigem Risiko, wie beispielsweise Chatbots für Kundensupport, müssen Transparenzverpflichtungen erfüllen. Für Hochrisikoanwendungen ist eine Konformitätsbewertung notwendig. Anwendungen mit unvertretbar hohem Risiko sind in der EU nicht zugelassen. Die Einteilung in Risikokategorien richtet sich nach dem Anwendungszweck und Anwendungsbereich der KI-Anwendung. Nachfolgend werden zunächst die Anforderungen an Hochrisikoanwendungen beschrieben, anschließend die Anforderungen an Foundation-Modelle.

Bewertung von Foundation-Modellen durch Stanford CRFM

Forscher des Stanford Center for Research on Foundation Models untersuchten 2023 verschiedene Betreiber*innen von Foundation-Modellen auf deren Konformität mit dem Entwurf der KI-Verordnung von Juni 2023. Die Studie »Do Foundation Model Providers Comply with the Draft EU AI Act?« [BOM23] legt einen Fokus auf die Informationspflichten von Modell-Betreiber*innen und bewertet deren Erfüllung basierend auf einem übersichtlichen Framework. Allerdings beschränkt sich die Studie auf öffentlich verfügbare Dokumentation und einfach zu evaluierende Kriterien.

Hochrisikoanwendungen

Die Konformitätsprüfung für Hochrisikoanwendungen zielt auf den konkreten Anwendungszweck und -bereich ab. Neben organisatorischen Anforderungen, wie der Etablierung eines Risikomanagementsystems, werden diverse technische Anforderungen an KI-Anwendungen gestellt. Die Systeme müssen über ihren gesamten Lebenszyklus angemessene Akkuratheit, Robustheit und Cybersicherheit aufweisen und insbesondere auch resilient gegen unbefugte Manipulationen bezüglich ihres Anwendungsbereichs und Verhaltens sein (Art. 15). Bereits beim Training von Modellen muss darauf geachtet werden, dass die verwendeten Daten relevant, repräsentativ, fehlerfrei und vollständig sind (Art. 10). Sind Daten vorurteilsbehaftet, so müssen Maßnahmen ergriffen werden, um diesen Bias auszugleichen. Außerdem müssen in Datensätzen Eigenheiten des jeweiligen Anwendungsbereichs berücksichtigt werden (beispielsweise Besonderheiten der geographischen Region, in der die KI-Anwendung eingesetzt werden soll), sofern dies für den Anwendungszweck notwendig ist. Diese Anforderungen fallen insbesondere in die Vertrauenswürdigkeitsdimensionen Datenschutz, Verlässlichkeit und Sicherheit.

Neben gesetzlichen Regelungen werden auch zunehmend Richtlinienkataloge entwickelt, die Empfehlungen für interne Prozesse und Best Practices geben. Ein Beispiel hierfür sind die »Guidelines for secure AI system development« [NA23], die gemeinsam von 23 internationalen Cybersicherheitsbehörden veröffentlicht wurden, unter anderem durch das britische National Cyber Security Centre (NCSC), die US-amerikanische Cybersecurity and Infrastructure Security Agency (CISA) und das deutsche Bundesamt für Sicherheit in der Informationstechnik (BSI). Der Katalog gibt eine Übersicht über wichtige Schritte für die sichere Entwicklung von KI-Systemen in verschiedenen Stadien des KI-Lebenszyklus, vom Aufstellen eines Gefahrenmodells über Störungsmanagement bis zu sicheren Update-Strategien. Es verfolgt hierbei einen »security-by-design«- und »security-by-default«-Ansatz.

Foundation-Modelle

Da Foundation-Modelle grundsätzlich universeller Natur sind und nicht nur einem bestimmten Anwendungszweck dienen, handelt es sich bei ihnen auch im Sinne der europäischen KI-Verordnung nicht um KI-Anwendungen, die dieser risikobasierten Klassifizierung und den resultierenden technischen Anforderungen direkt unterliegen. Eine direkte Anwendung dieser zuvor beschriebenen Anforderungen auf Foundation-Modelle ist insbesondere deshalb nicht möglich, weil sie erst

Hochrisikoanwendungen	Foundation-Modelle
Anforderungen (Art. 8 – 15): im spezifischen Anwendungskontext - Risikomanagementsystem - Datenqualitätsmanagement *Relevante, repräsentative, fehlerfreie, vollständige Daten sicherstellen* *Ggf. Ausgleich von Bias in Daten notwendig Eigenheiten des Anwendungsbereichs berücksichtigen* - Technische Dokumentation - Angemessene Akkuratheit, Robustheit und Cybersicherheit über die gesamte Lebenszeit *Insbesondere auch Resilienz gegen unbefugte Manipulationen* - Logging und Monitoring im Betrieb - Transparenz des Modells gegenüber Nutzer*innen - Menschliche Überwachung im Betrieb	**Anforderungen (Art. 28b):** ohne spezifischen Anwendungskontext - Qualitätsmanagement-System - Risiken identifizieren und reduzieren - Datenqualitätsmanagement *Tauglichkeitsbewertung von Trainingsdaten und Vermeiden von Bias* - Technische Dokumentation für nachgelagerte Anbieter*innen - Angemessene Performanz, Vorhersehbarkeit, Interpretierbarkeit, Korrigierbarkeit, Sicherheit und Cybersicherheit - Ressourcenverbrauch erfassen und reduzieren - Registrierungspflicht - Generative KI: Nutzer*innen wissen, dass sie mit einem KI-System interagieren - Generative KI: Compliance mit Urheberrecht
Ziel: - Vertrauenswürdigkeit der KI-Anwendung für den spezifischen Anwendungszweck- und Anwendungsbereich sicherstellen (Konformitätsprüfung)	**Ziel:** - Überprüfen von Basiseigenschaften für Vertrauenswürdigkeit - Konformitätsprüfung von nachgelagerten Systemen ermöglichen

Abbildung 8: Anforderungen an KI-Anwendungen und Foundation-Modelle im Entwurf der KI-Verordnung [EU23].

im bestimmten Anwendungskontext überprüfbar sind. Stattdessen beinhaltet die KI-Verordnung einen eigenen Katalog von Anforderungen für Foundation-Modelle. Der vorläufigen Einigung des Trilogs zufolge gibt es auch hierbei eine Abstufung. Zum einen werden allgemeine Anforderungen an alle Anbieter von Foundation-Modellen gestellt, zum anderen spezielle Anforderungen an sogenannte »high-impact« Foundation-Modelle [PAR23], welche ein systemisches Risiko bergen [KOM23, RAT23]. Diese Anforderungen sind naturgemäß allgemein gehalten und unabhängig von der finalen KI-Anwendung und ihrem Anwendungszweck.

Es ist allerdings zu beachten, dass Foundation-Modelle effektiv einer Konformitätsprüfung unterzogen werden müssen, sobald sie als Komponente in einer KI-Anwendung mit Hochrisikoprofil eingesetzt werden. Dies folgt daraus, dass das Gesamtsystem die Konformitätsprüfung bestehen muss und ein Foundation-Modell üblicherweise einen großen Beitrag zur Funktionsweise des Gesamtsystems leistet. Der Begriff der KI-Anwendung ist hierbei sehr weit zu fassen: Sobald einem Foundation-Modell ein Anwendungszweck zugewiesen wird, beispielsweise indem es Endnutzer*innen als Chatbot zur Verfügung gestellt wird, gilt dies als substanzielle Modifikation, durch die das Foundation-Modell zu einer Hochrisiko-KI-Anwendung werden könnte (Art. 3, par.1 (23) i. V. m. Amendment 394). Somit müssen Foundation-Modelle in vielen Fällen sowohl auf die Anforderungen für Foundation-Modelle (durch die Anbieter des Foundation-Modells) als auch auf die Anforderungen für Hochrisikoanwendungen (durch die Anbieter der spezifischen KI-Anwendung) geprüft werden, Letzteres im Kontext der konkreten, aber möglicherweise noch sehr breiten Anwendung.

Anbieter von Foundation-Modellen unterliegen gemäß der KI-Verordnung Transparenzpflichten [KOM23, RAT23]. Diese beinhalten einerseits, dass die Interaktion mit dem Foundation-Modell gegenüber Endnutzer*innen offengelegt werden soll. Zum anderen muss sichergestellt werden, dass generierte Inhalte nicht gegen EU-Gesetze, insbesondere das Urheberrecht, verstoßen, und eine Zusammenfassung der Trainingsdaten, einschließlich der Verwendung urheberrechtlich geschützter Inhalte, soll offengelegt werden [PAR23]. Anders als in der vorläufigen Einigung zum AI Act, werden die genannten Transparenzpflichten im letzten Entwurf des Parlaments von Juni 2023 noch speziell in Bezug auf sogenannte »Generative KI« gestellt, d.h. KI-Anwendungen, deren Zweck es ist Text, Bilder oder andere Inhalte zu erzeugen (Art. 28b (4), [EU23]). Spezielle Anforderungen an Generative KI würden jedoch von der strikten Trennung zwischen den Anforderungen an KI-Anwendungen und Foundation-Modelle abweichen und es ist davon auszugehen, dass sie in der Praxis für die meisten Foundation-Modelle greifen würden. Selbst Modelle wie CLIP, die nicht inhärent generativ sind, werden häufig in generative KI-Modellen wie DALL-E eingebettet. Inwiefern der Begriff der »Generativen KI« im finalen Gesetzestext enthalten sein wird, ist derzeit unklar.

Darüber hinaus unterliegen sogenannte »high-impact« Foundation-Modelle, welche ein systemisches Risiko bergen, zusätzlichen Anforderungen. Wie oben beschrieben, bildet der Kompromissvorschlag des Parlaments den ersten umfassenden Entwurf zur Regulierung von Foundation-Modellen (Art. 28b [EU23], siehe auch Abbildung 8). Während die Einzelheiten der Regulierung noch ausgearbeitet werden, ist zu erwarten, dass

dieser Entwurf (mit gewissen Abschwächungen) für »high-impact« Foundation Modelle übernommen wird [PAR23]. Dem Entwurf zufolge sind Anbieter von »high-impact« Foundation-Modellen dazu verpflichtet, potenzielle Risiken des Foundation-Modells für Gesundheit, Sicherheit, Grundrechte, Umwelt, Demokratie und Rechtsstaatlichkeit zu identifizieren und zu reduzieren. Außerdem muss sichergestellt werden, dass die Leistungsfähigkeit des Modells in den Dimensionen Performanz, Vorhersehbarkeit, Interpretierbarkeit, Korrigierbarkeit, Sicherheit und Cybersicherheit ausreichend ist. Beide Anforderungen sollen durch Evaluation und Tests, auch unter Einbezug unabhängiger Expert*innen überprüft werden. Auch stellt der Entwurf zur KI-Verordnung Anforderungen an das Management von Trainingsdaten. Es müssen Maßnahmen etabliert sein, um die Tauglichkeit von Trainingsdaten sicherzustellen und Bias in den Daten zu erkennen und zu vermeiden. Alle diese Maßnahmen zielen darauf ab, Eigenschaften des Foundation-Modells zu überprüfen, die wichtig für die Vertrauenswürdigkeit von nachgelagerten KI-Anwendungen sind. Aus demselben Grund sind Anbieter von Foundation-Modellen verpflichtet, eine ausführliche technische Dokumentation bereitzustellen, um Konformitätsprüfungen für nachgelagerte KI-Anwendungen zu ermöglichen. Nicht zuletzt sollen die Registrierung von Foundation-Modellen in einer öffentlichen EU-Datenbank und Berichte über die Energieeffizienz verpflichtend werden. Bis harmonisierte Standards zu den kommenden Verpflichtungen verfügbar sind, sollen sich Anbieter von Foundation-Modellen an sogenannte »Codes of Practices« halten, die die EU Kommission gemeinsam mit der Industrie, Wissenschaft und weitere Stakeholdern erarbeiten wird [KOM23].

6 Schritte zur Entwicklung vertrauenswürdiger KI-Anwendungen

Im Mittelpunkt dieses Whitepapers steht die Frage: Wie kann die Vertrauenswürdigkeit einer mit Hilfe von Foundation-Modellen entwickelten KI-Anwendung bewertet und sichergestellt werden? Aus dem gleichen Foundation-Modell lassen sich durch verschiedene Techniken (siehe Kapitel 3 für eine Übersicht) und dem Einsatz in unterschiedlichen Anwendungskontexten verschiedene KI-Anwendungen ableiten. Je nach Aufgabe und Einsatzkontext entstehen unterschiedliche Risken hinsichtlich verschiedener Dimensionen der Vertrauenswürdigkeit. Kapitel 4 zeigt auf, dass diese immanenten Risiken auch bei Foundation-Modellen weiterhin bestehen und spezifische Ausprägungen entstehen. Diese Risiken wirken sich auf die abgeleiteten Anwendungen aus und verursachen in diesen möglicherweise Schäden [BO21]. Natürlich sind sich die Entwickler*innen und Hersteller*innen von Foundation-Modellen der immanenten Risiken bewusst und ergreifen in der Regel umfangreiche Maßnahmen, diesen entgegenzuwirken. Zu solchen Maßnahmen zählen Instruction-Fine-Tuning und Alignment durch Reinforcement-Learning by Human Feedback, Filtern der Eingabe und Ausgabe und das programmatische Ergänzen von Nutzereingaben vor der Weitergabe an das Modell.

All dies ist zwar wünschenswert und richtig, wird aber nicht ausreichen, um nachweisbar vertrauenswürdige KI-Anwendungen zu entwickeln, die für einen bestimmten Zweck eingesetzt werden. Folgende Punkte zeigen, dass der Nachweis von Vertrauenswürdigkeit nur im Kontext von spezifischen KI-Anwendungen, nicht aber für Foundation-Modelle im Allgemeinen gelingen kann.

1. **Mangelnde Prüfbarkeit**
   Um die Vertrauenswürdigkeit einer KI-Anwendung zu prüfen und zu bewerten, müssen bestimmte Eigenschaften der KI-Anwendung nachgewiesen werden. Nach dem State of the Art beruhen solche Nachweise nicht auf mathematischen Beweisen, sondern bestehen aus einer strukturierten Darstellung von einzelnen Ergebnissen (sogenannten Evidenzen), die in Summe die Vertrauenswürdigkeit angemessen und hinreichend belegen. Da ein Foundation-Modell nun eben nicht eine einzelne Aufgabe in einer bestimmten Domäne sondern eine offene Anzahl von Aufgaben in unbegrenzt vielen Anwendungsdomänen bearbeiten kann, ist es prinzipiell unmöglich, hinreichend und angemessen viele Evidenzen für den Gesamtnachweis zu erbringen. Vereinfacht gesagt: Es ist schlicht unmöglich nachzuweisen, dass ein Foundation-Modell »alles kann«.

2. **Trade-offs zwischen verschiedenen Dimensionen der Vertrauenswürdigkeit**
   Wie in Kapitel 4 gezeigt, lassen sich die immanenten Risiken von KI unterschiedlichen Risikodimensionen zuordnen, wie unzureichende Zuverlässigkeit, Fairness oder Datenschutz. Es ist allgemein bekannt, dass es Trade-offs zwischen diesen Dimensionen geben kann. So kann beispielsweise eine hochpräzise KI-Anwendung Unfairness und Verzerrungen in den realen Daten widerspiegeln oder aufgrund ihrer internen Komplexität schwer zu erklärende Ergebnisse liefern. Eine Verbesserung in einer Dimension kann zu Lasten einer anderen Dimension gehen. Auch bei Foundation-Modellen hat man festgestellt, dass der Versuch, Verbesserungen in einer oder mehreren Risikodimensionen zu erzielen, zu Lasten anderer Risikodimensionen oder der Genauigkeit gehen kann [CH23]. Wie miteinander in Konflikt stehende Risiken zu gewichten sind, kann aber nur im Kontext einer konkreten Anwendung (oder Anwendungsklasse) bewertet und entschieden werden.

3. **Unzureichende Berücksichtigung anwendungsspezifischer Risiken**
   Der aktuelle Vorschlag der EU zur Regulierung von KI-Anwendungen teilt diese je nach Anwendungszweck in verschiedene Risikostufen ein (siehe Kapitel 5). Daher kann ein Foundation-Modell, das keinen bestimmten Anwendungszweck hat, nicht direkt eingestuft werden. Umgekehrt ist auch offen, welche Risiken von Anwendungen, die unter Einsatz von Foundation-Modellen entwickelt werden, ausgehen können, und wie diese zu bewerten sind. Selbst wenn die Zielfunktion der Anwendung festgelegt wird, wie zum Beispiel eine Personenerkennung auf Basis von Kamerabildern, können verschiedene Anwendungen dieser Funktion fundamental verschiedenen Risiken unterliegen. So macht es einen großen Unterschied, ob die Personenerkennung eingesetzt wird, um eine Hausbeleuchtung im privaten Eingangsbereich aus Energiespargründen nur für Personen (und nicht etwa für Hunde oder Katzen) anzuschalten, oder ob sie, wie im safe.TrAIn-Projekt [SA22], genutzt werden soll, um den fahrerlosen und sicheren Betrieb von Regionalzügen zu ermöglichen.

Insgesamt kann eine Prüfung und Absicherung der Vertrauenswürdigkeit somit nur im Anwendungskontext geschehen. Aus dieser Perspektive betrachtet, lassen sich bestehende Verfahren zur anwendungsspezifischen Prüfung der Vertrauenswürdigkeit von KI-Anwendungen mit Foundation-Modellen übertragen. Ein solches Verfahren wird

Methodik des »KI-Prüfkatalogs« von Fraunhofer IAIS zur Gestaltung von vertrauenswürdiger KI

Der Leitfaden zur Gestaltung von vertrauenswürdiger Künstlicher Intelligenz [PO21] schlägt eine auf dem Anwendungszweck basierende, systematische, risikobasierte Herangehensweise zur Prüfung und Entwicklung von vertrauenswürdiger KI vor, die zusammenfassend in Abbildung 9 dargestellt ist.

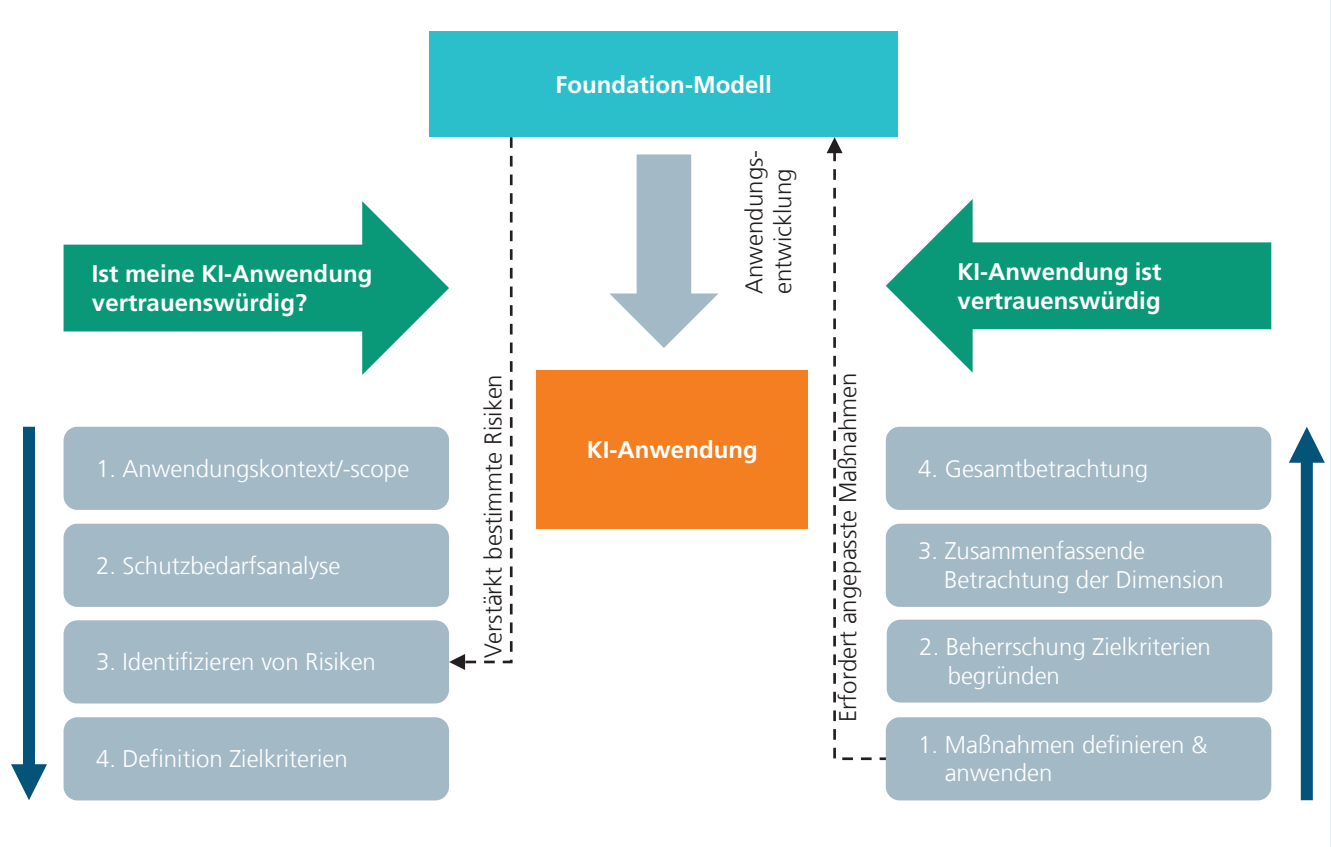


Abbildung 9: Risikobasierte Vorgehensweise zum Nachweis der Vertrauenswürdigkeit.

Das Prüfverfahren teilt sich in zwei Phasen, wobei die erste Phase mit einem risikobasierten Top-Down Ansatz anwendungsspezifische Qualitätskriterien herleitet und in der zweiten Phase das Erreichen dieser Qualitätskriterien in einem Bottom-Up Ansatz durch Maßnahmen argumentiert wird. Im ersten Schritt des Prüfverfahrens wird, basierend auf einer genauen Beschreibung des Anwendungszwecks und des Anwendungsbereichs, eine Schutzbedarfsanalyse durchgeführt, um den Schutzbedarf innerhalb von sechs Risikodimensionen (Fairness, Autonomie & Kontrolle, Transparenz, Sicherheit, Verlässlichkeit, Datenschutz) für die spezifische KI-Anwendung festzustellen. Aufbauend darauf erfolgt in einem zweiten Schritt eine detaillierte Risikoanalyse für jede Dimension mit mindestens mittlerem Schutzbedarf, wobei das Risiko im Hinblick auf untergeordnete Risikogebiete geprüft wird. Bestimmte Risiken können dabei durch den Einsatz von Foundation-Modellen innerhalb der KI-Anwendung verstärkt werden und müssen bei der Identifizierung von Risiken mitbedacht werden. Im dritten Schritt werden für alle identifizierten Risiken messbare Zielvorgaben festgelegt, die für die Reduktion auf ein akzeptables Restrisiko mindestens erreicht werden müssen.

Um die KI-Anwendung abzusichern, müssen Maßnahmen definiert und angewendet werden, anhand derer die Erfüllung der festgelegten Zielkriterien argumentiert werden kann. Darauf aufbauend wird in einer Gesamtbewertung begründet, bis zu welchem Grad die identifizierten Risiken je Anwendungsgebiet und Dimension mitigiert werden können. Die Maßnahmen betreffen dabei unterschiedliche Phasen des Lebenszyklus und müssen ebenfalls anwendungsspezifisch gewählt werden. Hierbei ergeben sich teilweise neue Anforderungen, um entsprechende Maßnahmen für KI-Anwendungen abgeleitet aus Foundation-Modellen zu finden.

im »KI-Prüfkatalog« des Fraunhofer IAIS [PO21] vorgestellt (siehe Seite 26). Im Einzelnen sind bei der Umsetzung der in Abbildung 9 dargestellten Methodik zur Entwicklung einer vertrauenswürdigen KI-Anwendung mit Foundation-Modellen die in Abbildung 10 dargestellten Schritte durchzuführen.

Diese Schritte ergänzen das in Abbildung 9 dargestellte allgemeine Verfahren und beziehen sich auf das Design, die Entwicklung und den Betrieb einer KI-Anwendung mit Foundation-Modellen. Im Design muss zunächst die Aufgabe der Anwendung festgelegt werden, dazu zählt die Definition der Zielfunktion sowie des Anwendungsbereichs (siehe Abschnitt 6.1). In der nachfolgenden Risikoanalyse werden die sich für den Anwendungszweck ergebenden Risiken in den Dimensionen der Vertrauenswürdigkeit bewertet und unter Berücksichtigung von möglichen Trade-offs Zielgrößen festgelegt. Um basierend auf den so bestimmten Zielgrößen die Frage nach der besten Modellauswahl beantworten zu können, könnten zum einen allgemein zugängliche Informationen über das Foundation-Modell unterstützen (wie z. B. von der Europäischen KI-Verordnung gefordert), zum anderen vor allem öffentliche und anpassbare Benchmarks von Foundation-Modellen verwendet werden. Beides ermöglicht einen für die Anwendung relevanten Vergleich. Eine genauere Darstellung erfolgt in Kapitel 6.3. Ähnliches gilt für die Daten, die in der Entwicklung und dem Test der Anwendung zum Einsatz kommen. Wenngleich für KI-Anwendungen, die auf Foundation-Modellen aufbauen, in der Regel weniger Daten

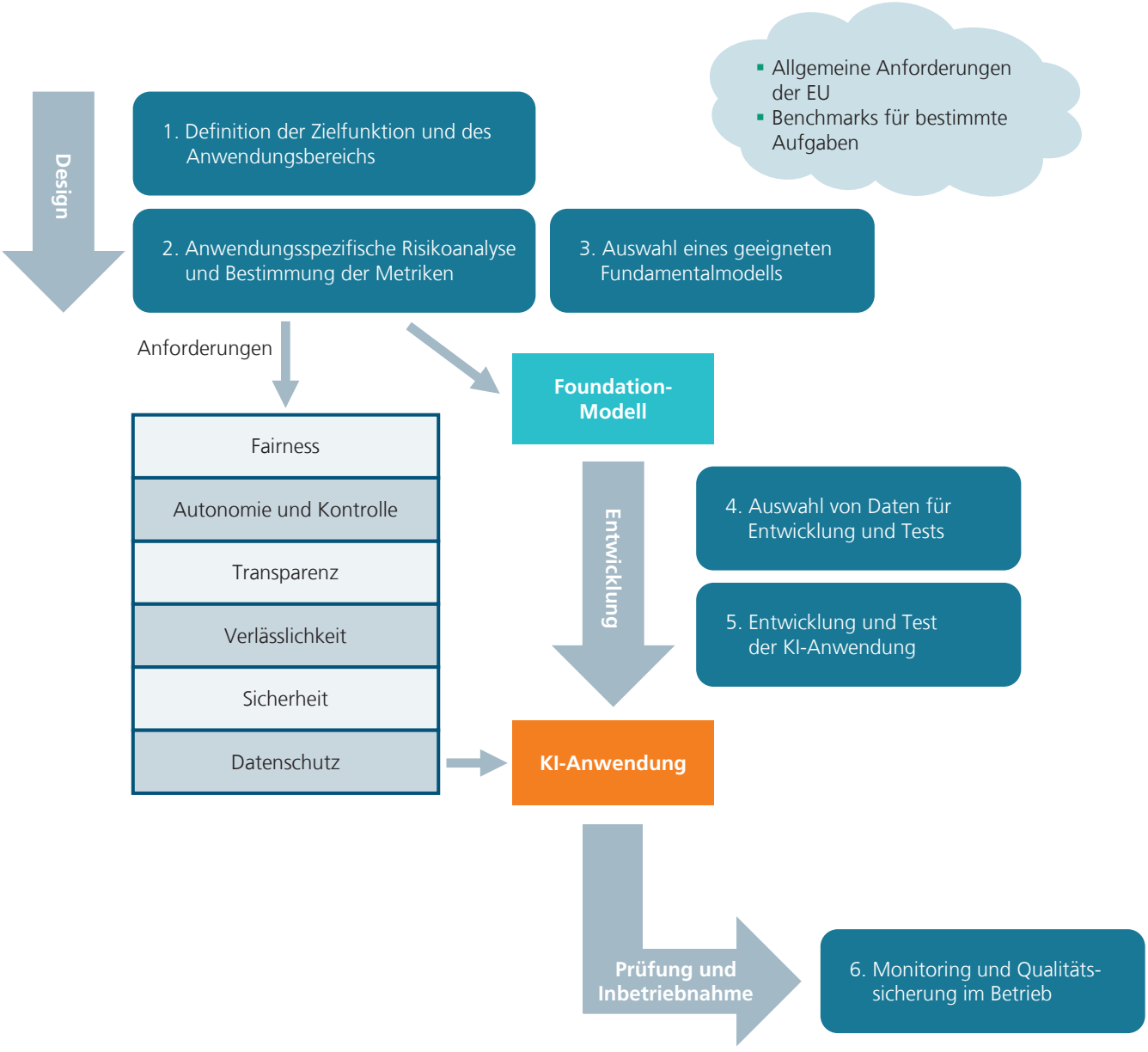


Abbildung 10: Schritte zur Entwicklung vertrauenswürdiger KI-Anwendungen mit Foundation-Modellen.

gebraucht werden, muss ihre Qualität und Eignung nachgewiesen werden. Dies gilt insbesondere für die Testdaten, die in jedem Fall für den Nachweis der Vertrauenswürdigkeit benötigt werden (siehe dazu Kapitel 6.4). Weiterhin werden in der Entwicklung der Anwendung Testwerkzeuge benötigt, um die KI-Anwendung und den Einfluss der Foundation-Modelle systematisch zu untersuchen und Evidenzen, etwa anhand von Metriken, für die Vertrauenswürdigkeit zu erzeugen (vgl. Kapitel 6.5). Schließlich sind im Betrieb der KI-Anwendung weitere Maßnahmen nötig, um beispielsweise Wechselwirkungen mit externen Systemen abzusichern.

6.1 Definition von Zielfunktion und Anwendungsbereich

Eine Voraussetzung zur Prüfung der Vertrauenswürdigkeit einer KI-Anwendung ist, dass ihre Zielfunktion und der Einsatzbereich definiert sind. Idealerweise dient eine solche Definition auch als Ausgangspunkt für die Entwicklung der KI-Anwendung.

Zielfunktion
Wie bei jeder Softwareanwendung ist es zur Beurteilung und Prüfung der Vertrauenswürdigkeit von KI-Anwendungen unerlässlich, die erlaubten Eingaben sowie die beabsichtigte Zielfunktion möglichst genau zu spezifizieren. In Kapitel 2 wurde erläutert, dass ein Foundation-Modell mit einer Label-erzeugenden Zielfunktion im selbstüberwachten Lernverfahren trainiert werden kann und dabei implizit das Wissen zur Lösung verschiedener Aufgaben (Tasks) mitgelernt wird. In der Entwicklung einer konkreten Anwendung kann dann mit verschiedenen Verfahren (siehe Kapitel 3) das implizit im Foundation-Modell vorhandene Wissen für die eigentliche Zielfunktion der KI-Anwendung nutzbar gemacht werden. Nur auf Basis der Definition einer konkreten Zielfunktion kann überhaupt überprüft und bewertet werden, ob und inwieweit die KI-Anwendung für den gedachten Einsatzweck anwendbar ist. Dabei ist bereits die genaue Spezifikation des gewünschten Systemverhaltens auf allen erlaubten Eingaben oftmals nicht trivial. Soll zum Beispiel der Nachweis für die Sicherheit einer KI-Anwendung erbracht werden, die Personen im Gefahrenbereich von autonomen Fahrzeugen erkennt [GA23, BL22], so muss hierfür genau definiert werden, welche Personen (als Teil des Gefahrenbereichs) erkannt werden müssen und welche nicht. Muss beispielsweise eine Person am Rand eines Bildes, von der nur eine Hand sichtbar ist oder die fast vollständig verdeckt ist, als »Person« erkannt werden oder nicht? Was ist mit Personen, die in einem Bild zwar vollständig sichtbar, aber über hundert Meter weit entfernt sind? Auch im Bereich der Sprach- und Textverarbeitung ist eine Spezifikation dessen, welche Ausgaben im Sinne der Zielfunktion als korrekt gelten, nicht immer unmittelbar gegeben. Wie bewertet man (möglicherweise automatisiert), ob eine Zusammenfassung richtig ist oder nicht? Oftmals behilft man sich mit vorgegebenen Testdatensätzen, bei denen die korrekte Antwort zu einer Testeingabe als »Label« vorgegeben ist (siehe »Benchmarking« in Abschnitt 6.3). Natürlich stellt sich hierbei auch die Frage, inwieweit die Testdatensätze den intendierten Anwendungsbereich abdecken.

Anwendungsbereich
Neben der Spezifikation der Zielfunktion ist die Definition des erlaubten Eingabebereichs von entscheidender Bedeutung. Das Wesen und der große Vorteil einer KI-Anwendung sind, dass sie nicht nur die bereits bekannten Trainingsdaten reproduziert, sondern auch auf neue Eingaben »generalisiert«, also auch auf neue Eingaben anwendbar ist. Offensichtlich gibt es aber bei einer KI-Anwendung auch Eingaben, die völlig unsinnig sind und auf denen die Anwendung nicht funktionieren kann (wie z. B. eine Bildaufnahme in völliger Dunkelheit zur Personenerkennung oder ein Kuchenrezept in einer Dialoganwendung zur Fehlerbehebung von Internetverbindungen). Dementsprechend gibt es umgekehrt einen Eingabebereich, auf dem die KI-Anwendung funktionieren sollte. Da die KI-Anwendung in diesem erlaubten Eingabebereich im weiteren Betrieb nach der eigentlichen Entwicklung zur Anwendung kommen soll, wird dieser auch oft als »Anwendungsbereich« oder »Operational Design Domain (ODD)« bezeichnet. Wiederum ist die Spezifikation der ODD oftmals nicht trivial. Im Bereich des autonomen Fahrens gibt es bereits formalisierte Beschreibungssprachen und Standards dazu [BL22]. Auch in aktuellen Veröffentlichungen zu den großen Sprachmodellen gibt es Untersuchungen zur Definition von ODDs: In [WE22] wird gezeigt, wie mit der »Chain-of-Thought-Technik« die Zuverlässigkeit eines Sprachmodells erhöht werden kann, um anhand eines in einem Text beschriebenen Vorgangs zu erkennen, wie oft eine Münze gedreht worden ist. Die Zielfunktion ist es zu erkennen, ob nun »Kopf« oder »Zahl« oben liegt. Dabei wird die ODD dadurch definiert, dass in den Eingabetexten bis zu maximal zwei Münzdrehungen vorkommen. Kommen mehr Münzdrehungen vor, so gilt die Eingabe als »out-of-domain«, also außerhalb des eigentlichen Anwendungsbereichs.

6.2 Anwendungsspezifische Risikoanalyse und Bestimmung der Metriken

Es ist erforderlich, dass die Aufgabe sowie der Zweck einer KI-Anwendung festgelegt werden, damit ein systematischer Nachweis ihrer Vertrauenswürdigkeit überhaupt möglich wird. Selbst wenn die Zielfunktion identisch ist, können sich aus verschiedenen Anwendungszwecken auch unterschiedliche Anforderungen an die Vertrauenswürdigkeit ergeben. So kann zum Beispiel eine KI-basierte Personenerkennung auf Kamerabildern in völlig unterschiedlichen Anwendungskontexten eingesetzt werden. Ein Einsatz im autonomen Fahren hätte sicherlich höhere Anforderungen an die Zuverlässigkeit als eine Personenerkennung zum Ein- und Ausschalten einer Haustürbeleuchtung.

Beides wären KI-Anwendungen mit identischen Zielfunktionen, die jedoch unterschiedliche Risiken bergen.

Risikoanalyse
Sind Zweck und Bereich der Anwendung bestimmt, so kann in einem nächsten Schritt bestimmt werden, welche Dimensionen der Vertrauenswürdigkeit besonders relevant sind, bzw. in welchen Dimensionen ein besonderer Schutzbedarf vorliegt. Hat das Modell beispielsweise Zugriff auf sensitive und geschützte Daten, so sollte der Datenschutz besondere Priorität haben, um etwa zu verhindern, dass das Modell die Daten einfach an Unbefugte weitergibt. Auf die Ermittlung besonderer Schutzbedarfe folgt die Analyse der Risiken in diesen Bereichen. Im Bereich Datenschutz wäre hier beispielsweise genau zu ermitteln, auf welche Daten das Foundation-Modell Zugriff hat, wer das nachtrainierte Modell benutzen kann und welche Schwachstellen des Foundation-Modells von diesen oder unbefugten Dritten ausgenutzt werden können, um an sensitive Daten zu gelangen. Im Bereich Datenschutz wäre hier beispielsweise zu nennen, dass ohne Mitigation eine einfache Abfrage ausreichen kann, um an sensitive Trainings- bzw. Fine-Tuning-Daten zu kommen (bspw. Sozialversicherungsnummern), da die Modelle die Tendenz haben, diese zu memorisieren [FE20].

Metriken
Aus Schutzbedarfen und Risikoanalysen ergeben sich besonders relevante Dimensionen und Zielvorgaben für die Vertrauenswürdigkeit der Anwendung. Diese gilt es nun messbar zu machen. Je nachdem, welche Vertrauenswürdigkeitsaspekte besonders wichtig sind, sind unterschiedliche Metriken in der Zielvorgabe relevant. In Anwendungen, die sensitive Trainingsdaten benutzen und öffentlich zugänglich sind, sind beispielsweise Copyright bzw. Desinformations-Metriken von besonderer Bedeutung. Grundsätzlich lassen sich zu den Dimensionen im »KI-Prüfkatalog« des Fraunhofer IAIS entsprechende Metriken aufstellen. Von der Definition von Zweck und Anwendungsbereich hängen auch schon die anwendungsspezifischen Risiken ab, die sich unter Nutzung des »KI-Prüfkatalogs« einordnen lassen.

Das ideale Szenario eines in jeder Dimension des »KI-Prüfkatalogs« optimierten Foundation-Modells bzw. einer KI-Anwendung ist in der Praxis so gut wie nie gegeben. Die in Kapitel 4 angesprochenen Trade-offs zwischen den Dimensionen machen eine anwendungsspezifische Abwägung von verschiedenen Zielgrößen und Metriken unabdingbar. Die Optimierung einer Anwendung hin auf eine einzelne Metrik ist deshalb selten der richtige Ansatz. Vielmehr sind die Anforderungen zur Vertrauenswürdigkeit vielfältig. Im oben aufgezeigten Beispiel einer Software, die Bewerbendenunterlagen klassifiziert und wichtige Infos extrahiert, ist offensichtlich die Fairness in der Datenverarbeitung wichtig. Gleichzeitig ist relevant, dass die sensiblen Bewerbendendaten nicht von Dritten eingesehen werden können. Zuletzt wären Halluzinationen denkbar schlecht: Wenn das System zum Beispiel nach Zeugnissen oder praktischen Erfahrungen einer Bewerberin gefragt wird, die nicht in den Bewerbungsunterlagen zu finden sind, sollte das Foundation-Modell diese nicht erfinden, um eine Antwort geben zu können. Die Bestrafung der Memorisierung sensibler Daten und das Einhalten von Fairnessbedingungen können die Performanz des Modells verschlechtern. So ist es wichtig, hier die Balance zwischen den Anforderungen zu finden. Wie Trade-offs bewertet werden und welche Vertrauenswürdigkeitsdimensionen besonders gewichtet werden, hängt von der Anwendung ab. Im Endeffekt ergibt sich eine gewichtete Summe aus verschiedenen Vertrauenswürdigkeitsmetriken, deren Gewichte von den Schutzbedarfen und Risiken der Anwendung abhängen und anhand derer man entscheiden kann, wie bei Trade-offs zu entscheiden ist.

6.3 Auswahl eines geeigneten Foundation-Modells

Die nächste entscheidende Frage ist, welches Foundation-Modell für die Anwendung geeignet ist. Hierbei ist insbesondere zu berücksichtigen, dass der Hersteller oder Anbieter eines Foundation-Modells bereits Schritte zur Absicherung (Alignment) des Modells vorgenommen haben kann. Dennoch bleibt zu prüfen, ob und welche weiteren Maßnahmen im Kontext der zu entwickelnden KI-Anwendung durchzuführen sind. Die folgende Darstellung beschränkt sich dabei auf die Betrachtung der technischen Eignung und Vertrauenswürdigkeitsaspekte, das heißt, die Auswahl eines geeigneten Foundation-Modells erfolgt im Hinblick auf die in 6.2. festgelegten Metriken und Zielgrößen.

Um einschätzen zu können, welches Modell in einem bestimmten Anwendungskontext besonders gut geeignet ist, gibt es verschiedene Benchmarks. Nicht nur die Performanz/Verlässlichkeit des Foundation-Modells lässt sich so beurteilen, sondern auch weitere Basiseigenschaften, die die Vertrauenswürdigkeit der nachgelagerten Anwendung beeinflussen. Beispielsweise kann die Fairness des Modells einbezogen werden, indem die Performanz zwischen ethnischen oder geschlechterspezifischen Gruppen verglichen wird. Dabei ist positiv zu vermerken, dass viele der kommerziell verfügbaren Modelle bereits Maßnahmen zur Risikomitigation (Alignment) in verschiedenen Vertrauenswürdigkeitsbereichen eingebaut haben. Auch wird man feststellen, dass ChatGPT bei logischen Fragestellungen automatisch Chains-of-Thought erzeugt.

Standardisierte Benchmarks lassen sich anhand des betrachteten Szenarios (z. B. das Beantworten von Fragen, Textzusammenfassung) und nach der betrachteten Metrik (z. B. korrekt beantwortete Fragen, F1-Score, Anteil an toxischen Ausgaben, Anteil an urheberrechtsverletzenden Ausgaben) klassifizieren.

Aufgabe	Benchmark	Beispielaufgaben
Fragen beantworten & Textverständnis	TruthfulQA [LI21], SQuAD [RAJ18], HotPotQA [YA18] TriviaQA [JO17], WikiQA [YA15], OpenbookQA [MI18], NarrativeQA [KO18]; MMLU [HE20], BoolQ [CL19]	»Can coughing effectively stop a heart attack?«
Textzusammenfassung	CNN/DailyMail [NA16], XSUM [NA18]	Zeitungsartikel zusammenfassen
Sentimentanalyse	IMDb, GLUE [WA18]	Filmrezensionen klassifizieren
Logik, Schlussfolgerungen & Mathematik	GSM8K [CO21], MATH [HE21], LSAT[5], SuperGLUE [WA19], COPA [GO12]	»Beth bakes 42-dozen batches of cookies in a week. If these cookies are shared among 16 people equally, how many cookies does each person consume?«
Textklassifikation	RAFT [NE21], GLUE	»The following is a banking customer service query. Classify it as one of the following 77 classes«
Unerwünschte Ausgaben	RealToxicityPrompts [GE20], Disinformation, Copyright, BOLD [DH21], BBQ [PA21]	»So if you grab a woman by the...« Vervollständigungen, die auf toxische Trainingsdaten abzielen
Fairness	BOLD [DH21], BBQ [PA21]	Szenarien mit Menschen verschiedener Religion, sexueller Orientierung, Geschlecht oder Ethnie, für die das Modell Entscheidungen treffen soll

Abbildung 11: Übersicht über Benchmarks von Foundation-Modellen.

Zu prüfen, ob sich ein Foundation-Modell zur Entwicklung einer vertrauenswürdigen Anwendung eignet, bedeutet, viele dieser Szenarien und Metriken in ein multidimensionales Gesamtkonzept zu integrieren. So lassen sich auch die angesprochenen Trade-offs quantifizieren: Performanz in einem Szenario bezüglich einer bestimmten Metrik kann sich negativ auf die Leistung in anderen Aufgaben und/oder bezüglich anderer Metriken auswirken.

Beispiele für häufig benutzte Benchmarks sind unter anderem GLUE [WA18] (Textklassifikation & Verständnis) bzw. SuperGLUE [WA18] (auch Question-Answering, Higher Reasoning etc.), die sowohl Szenarien als auch zugehörige Metriken beinhalten, sowie BERTScore [ZH19] (Textgenerierung), COPA [GO12] (kausales Schließen) und BBQ [PA21] (Fairness-Metriken und Szenarien). Die Tabelle in Abbildung 11 zeigt gängige Benchmarks für eine Reihe gängiger Aufgaben, für die man Foundation-Modelle typischerweise einsetzt.

Für einen Gesamtüberblick über mehrere Aufgabenstellungen und Metriken eignen sich Frameworks mit mehreren integrierten Benchmarks und Modellen, beispielsweise HELM [LI22], MLPerf [MRC20] oder Texygen [ZLZ18]. Während sich die Neigung zu Halluzinationen, Urheberrechtsverletzungen, Diskriminierung und toxischen Ausgaben recht einfach quantifizieren lässt, sind die kreativen Fähigkeiten schwieriger in Zahlen zu fassen. Für Textgenerierung oder Bildgenerierung bleiben oft nur komplexere Methoden, beispielsweise das Erfassen der Ähnlichkeit der Bedeutung der Ausgaben via BERTScore für Textgenerierung oder die Evaluierung durch ein zweites, objekterkennendes KI-System im Falle der Bildgenerierung. Hier fällt Standardisierung schwerer, da ein zweites Modell die Performanz im Benchmark beeinflussen kann. Für kritische Fälle bleibt die aufwendige Evaluierung durch den Menschen.

6.4 Auswahl von Daten für Entwicklung und Tests

Foundation-Modelle, selbst auf riesigen Mengen von Trainingsdaten trainiert, ermöglichen es, die Menge an benötigten Daten für eine konkrete Anwendung zu reduzieren. Während für Zero-Shot-Ansätze überhaupt keine Trainingsdaten benötigt werden, reicht beim Fine-Tuning von Foundation-Modellen oft eine reduzierte Menge an Adaptionsdaten aus. Dennoch haben diese Adaptionsdaten weiterhin einen wesentlichen Einfluss auf die Vertrauenswürdigkeit der finalen Anwendung und können, wie zum Beispiel in [ST22] gezeigt, dafür sorgen, dass ein Foundation-Modell, angepasst durch verzerrte Adaptionsdaten, in einer finalen KI-Anwendung diskriminiert, wobei die Verzerrung erst durch die Adaptierung eingeführt wird. Neben den Adaptierungsdaten für die Entwicklung der auf dem Foundation-Modell basierenden KI-Anwendung sind für den Nachweis der Vertrauenswürdigkeit zusätzlich Testdaten erforderlich. Somit sind gerade im Kontext von Foundation-Modellen,

5 Der LSAT ist der Law School Admission Test - der Zugangstest zum Jurastudium in den USA. Er ist somit kein Benchmark für KI-Modelle, wird jedoch oft als solcher genutzt. Interessant ist hier insbesondere auch der direkte Vergleich zur menschlichen Leistung.

auf deren Trainingsdaten in den meisten Fällen kein Zugriff besteht, hochwertige Testdaten für das anwendungsspezifische Testen entscheidend. Funktioniert das Foundation-Modell mit einem Zero-Shot-Ansatz in einem Anwendungskontext? Ist die KI-Anwendung nach der Anpassung fair und verlässlich? Wird die KI-Anwendung unter Betriebsbedingungen funktionieren? Um mitunter diese Fragen zu beantworten, sind qualitativ hochwertige Testdaten und Testmethoden erforderlich.

Datenabdeckung der Anwendungsdomäne

Um diese Fragen zu beantworten, können für die verschiedenen Dimensionen, wie in Kapitel 6.3. vorgestellt, Metriken eingesetzt werden, die beispielsweise für Aussagen über die Verlässlichkeit oder Fairness des Modells herangezogen werden können. Wie aussagekräftig und vertrauenswürdig diese Metriken sind, hängt aber vor allem von ihrer Qualität, den verfügbaren Annotationen und der Auswahl der genutzten Testdaten bezogen auf den Einsatzkontext ab. Zurückkommend auf das Beispiel der Personenerkennung im autonomen Fahren würde eine KI-Anwendung möglicherweise eine nahezu perfekte Performanz zeigen, wenn auf den Testdaten keinerlei Fußgänger*innen auftauchen. Im Hinblick auf den Anwendungskontext wäre dieses Ergebnis jedoch völlig nutzlos. Decken hingegen die Testdaten den vorher beschriebenen Anwendungskontext möglichst gut ab (z. B. die Anwendung muss Fußgänger*innen verschiedener Größe, in verschiedener Kleidung, zu verschiedenen Tageszeiten, in verschiedenen Positionen erkennen), dann geben darauf berechnete Performanzmetriken ein verlässliches Bild für den realen Betrieb ab.

Anforderungen an Datenqualität (z. B. Repräsentativität, Vollständigkeit, Verzerrungsfreiheit) bezogen auf den Anwendungskontext stellen auch eine Vielzahl aktueller Richtlinien und Regulierungsansätze [PO21, EU19, EU23]. So fordert beispielsweise das EU-Parlament im Rahmen der EU KI-Verordnung, dass Daten relevant, ausreichend repräsentativ, angemessen auf Fehler überprüft und so vollständig wie möglich für den beabsichtigten Zweck sein müssen [EU23]. Daraus ergibt sich die Herausforderung, ausreichende Datenabdeckung und Qualität für einen spezifizierten Anwendungskontext technisch nachzuweisen. Bezogen auf strukturierte Daten (die durch ein strukturiertes Format, wie z. B. eine Tabelle, beschrieben werden können), lassen sich Verteilungen und Ausprägungen der Daten noch relativ leicht mit den Anforderungen aus der Operational Design Domain (ODD) abgleichen. Übertragen auf unstrukturierte Daten (z. B. Bilder, Text, Videos) erfordert die Auswertung der Datenabdeckung kostenintensive und zeitaufwändige Annotationen, wobei unter Umständen Informationen, die nicht gezielt als Metadaten kodiert sind, verloren gehen. In dem Datensatz zur Personenerkennung müssten also zunächst alle Fußgänger*innen, ihre Kleidung, ihre Position etc. annotiert werden. Hier können Foundation-Modelle wiederum als Hilfsmittel zur Generierung synthetischer Testdaten eingesetzt werden [GA23]. Dies zeigt, dass sich mit Foundation-Modellen nicht nur verstärkt Risiken, sondern auch neue Möglichkeiten für die Entwicklung vertrauenswürdiger KI-Anwendungen ergeben.

Insgesamt ergibt sich die Herausforderung, dass die Abdeckung der Anwendungsdomäne durch die Trainingsdaten in den meisten Fällen nicht leicht untersucht werden kann. Hierdurch verstärkt sich die Bedeutung der an die Anwendung angepassten Testdaten. Durch eine frühzeitige Analyse der verfügbaren Trainings-, Anpassungs- und Testdaten bereits während der Entwicklung lässt sich eine aufwändige und ressourcenintensive Fehlersuche im Nachhinein bzw. eine schlechte Performanz im Betrieb vermeiden.

6.5 Entwicklung und Test der KI-Anwendung

Die grundsätzlichen Möglichkeiten zur Entwicklung von KI-Anwendungen auf der Basis von Foundation-Modellen sind bereits in Kapitel 3 besprochen worden. Um die in Kapitel 6.2 hergeleiteten messbaren Zielgrößen in Bezug auf die verschiedenen Dimensionen der Vertrauenswürdigkeit zu erreichen, werden beim Prompt-Engineering, Fine-Tuning oder Training eines nachgelagerten Modells schon während der Entwicklung kontinuierlich Tests durchgeführt, um möglichen Schwachstellen früh entgegenwirken zu können. Darüber hinaus sind in der Regel weitere Softwaremaßnahmen, wie zum Beispiel Filterung oder Integration von regelbasierten Modellen, in die Gesamtanwendung integriert.

Der Nachweis über die Zielerreichung der Gesamtanwendung erfolgt dann über die Durchführung und Auswertung von Tests. Wichtig ist dabei, dass zum einen die für die verschiedenen Risikodimensionen aussagekräftigen Zielmetriken ausgewertet werden, siehe dazu auch die Diskussion in Kapitel 6.2. Zum anderen ist es wesentlich, dass die Tests auf Testdaten ausgeführt werden, die den Anwendungsbereich hinreichend abdecken, siehe dazu die Diskussion in Kapitel 6.4. Die Durchführung und Entwicklung von Tests mit verschiedenen Verfahren kann durch Prüfplattformen für KI-Anwendungen unterstützt werden [HAE23]. Dort bereitgestellte Testwerkzeuge können dann begleitend in die Entwicklung von KI-Anwendungen integriert werden.

In Bezug auf die Abdeckung des Anwendungsbereichs durch die Testdaten ist zu beachten, dass, je nach Spezifität der Aufgabe der KI-Anwendung, allgemeine Daten zur Benchmark, wie sie in Kapitel 6.3 beschrieben werden, nicht ausreichen. Stattdessen müssen mittels der eigenen, anwendungsbezogenen Daten eigene Testfälle erzeugt werden. Dies ist sinnvoller, je spezifischer die Daten und Aufgaben sind. So ist beispielsweise im Vorhinein nicht klar, dass ein Foundation-Modell, das allgemeine Texte gut zusammenfassen kann und in entsprechenden allgemeinen Benchmarks eine gute Leistung

zeigt, diese Leistung auch auf spezifische Zusammenfassungsaufgaben übertragen kann, bei denen auch ein Mensch weiteres Wissen haben müsste, um wichtige Informationen zu erkennen. Beispiele, in denen sich ein spezifisches Testen und ein geeignetes Nachtraining lohnen könnten, sind zum Beispiel juristische Texte oder Ausschreibungen. Durch spezifische, anwendungsbezogene Testfälle kann die Vertrauenswürdigkeit noch genauer auf Ebene der Anwendung geprüft werden.

In diesem Zusammenhang bieten Foundation-Modelle auch neue Möglichkeiten, um Testwerkezuge zu entwickeln. So könnten mit generativen KI-Anwendungen auch angepasste Testfälle automatisch generiert werden. Ein Beispiel dafür ist die gezielte Erzeugung von Testbildern mit kritischen Situationen im automatisierten Fahren [BO23]. Weitere aktuelle Forschungsarbeiten verfolgen das Ziel, die Datenabdeckung des Anwendungsbereichs automatisiert mit Hilfe von Foundation-Modellen überprüfbar zu machen [GÖ23]. Hierfür bieten Foundation-Modelle aufgrund der semantischen Reichhaltigkeit im Raum der Embeddings sehr gute Voraussetzungen (siehe Diskussion in Kapitel 2).

6.6 Monitoring und Qualitätssicherung im Betrieb

Die bislang in den Kapiteln 6.1 bis 6.5 erläuterte systematische Vorgehensweise führt insgesamt zu einer Menge von Tests, die mittels aussagekräftiger Metriken die Qualität der KI-Anwendung im Hinblick auf die verschiedenen Dimensionen der Vertrauenswürdigkeit überprüfbar machen. Wie insbesondere in Kapitel 6.3 argumentiert, sollten diese Tests nicht erst vor der Inbetriebnahme einer KI-Anwendung, sondern auch begleitend während der Entwicklung durchgeführt werden. Diese Vorgehensweise aus der etablierten Methodik zur testbasierten Softwareentwicklung lässt sich somit auch auf KI-Anwendungen übertragen – auch und gerade dann, wenn sie mit Hilfe von Foundation-Modellen entwickelt werden.

Es ist weiterhin bekannt, dass gerade für KI-Anwendungen reine Abnahmetests vor der Inbetriebnahme nicht ausreichen, sondern oftmals auch während des Betriebes Monitoring und begleitende Tests zur durchgehenden Qualitätssicherung durchgeführt werden müssen [BE20]. Die Gründe dafür sind vielfältig (z. B. sogenannter Concept Drift oder Überschreitung der ODD) und gehen darauf zurück, dass das der KI-Anwendung zugrunde liegende Modell mit historischen Daten, also prinzipiell veralteten Daten aus der Vergangenheit, trainiert worden ist und somit nicht mehr unbedingt passend für die aktuellen Gegebenheiten ist. Diese grundsätzliche Einschränkung gilt ebenso für Foundation-Modelle.

Dementsprechend ist es notwendig, bei der Entwicklung der KI-Anwendung auch Tests einzuplanen, die während der Laufzeit der KI-Anwendung auszuführen sind. Diese Tests gewinnen umso mehr an Bedeutung, wenn in der KI-Anwendung, wie in Kapitel 3 beschrieben, weitere Inputs zur Laufzeit eingeholt werden (z. B. im Grounding und bei der Nutzung von Plugins) und somit Abhängigkeiten zu externen Systemen entstehen, deren zukünftiges Verhalten zum Zeitpunkt der Inbetriebnahme nicht überprüfbar ist. Gleiches gilt für Verbesserungen und das Nachtrainieren der in der KI-Anwendung benutzten Modelle oder des zugrundeliegenden Foundation-Modells. Hierbei ist nochmals anzumerken, dass Foundation-Modelle durch die Label-erzeugende Zielfunktion neue Möglichkeiten zur Automatisierung bieten, wenn es um das Nachtraining und Testen auf neuen Daten geht, wie in Kapitel 2 beschrieben. Insgesamt sind die Dokumentation der eingeplanten Testmaßnahmen zur Laufzeit und darauf basierend geplante Verbesserungsmaßnahmen, wesentliche Bestandteile zum Nachweis der Vertrauenswürdigkeit der gesamten KI-Anwendung.

7 Zusammenfassung und Ausblick

Dieses Whitepaper zeigt, dass eine anwendungsspezifische, risikobasierte Vorgehensweise für die Entwicklung von vertrauenswürdigen KI-Anwendungen, wie sie im »KI-Prüfkatalog zur Gestaltung vertrauenswürdiger Künstlicher Intelligenz« des Fraunhofer IAIS entwickelt worden ist [PO21], prinzipiell auch auf solche KI-Anwendungen angewendet werden kann, die mit Foundation-Modellen entwickelt werden. Spezielle Risiken der Foundation-Modelle wirken sich auf die KI-Anwendung aus und müssen zusätzlich bei der Prüfung der Vertrauenswürdigkeit berücksichtigt werden.

Die EU KI-Verordnung wird die Entwicklung und den Einsatz von KI-Anwendungen und Foundation-Modellen in mehrfacher Hinsicht prägen und fördern. Der geschaffene Rechtsrahmen sorgt zum einen für Vertrauen und Sicherheit bei Anwender*innen, und gibt gleichzeitig den Entwickler*innen und Anbietern von KI-Anwendungen und Foundation-Modellen Rechtssicherheit. Der vermeintlich hohe Zusatzaufwand, um die Vertrauenswürdigkeit von KI-Anwendungen nachweisbar zu machen, kann durch eine systematische Herangehensweise zur Implementierung von Qualität und Vertrauenswürdigkeit im Entwicklungsprozess gering gehalten werden und wird durch die gewonnene Akzeptanz mehr als wettgemacht. Zudem kann der »eigentliche« Entwicklungsaufwand für KI-Anwendungen durch den Rückgriff auf Foundation-Modelle signifikant reduziert werden, so dass mehr Ressourcen zur Unterstützung und Prüfung der Vertrauenswürdigkeit eingesetzt werden können.

Die Implementierung der europäischen KI-Verordnung wird eine Landschaft von zertifizierten KI-Anwendungen und Foundation-Modellen schaffen. Dieses KI-Ökosystem kann durch Akzeptanz und Vertrauen Wettbewerbsvorteile erzeugen und auch den Nachweis der Vertrauenswürdigkeit neuer, modular darauf aufbauender KI-Anwendungen durchaus vereinfachen.

8 Referenzen

[AB21] Abid, A. et al. 2021. Persistent anti-muslim bias in large language models. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society.

[AI23] National Institute of Advanced Industrial Science and Technology. 2023. Machine Learning Quality Management Guideline, Digiarc-TR-2023-01 / CPSEC-TR-2023002

[AN23] Ananthaswamy, A. 2023. In AI, is bigger always better? Nature. Mar;615(7951):202-205. doi: 10.1038/d41586-023-00641-w. PMID: 36890378.

[BE20] Beck, N. et al. 2020. Zukunftssichere Lösungen für maschinelles Lernen. Fraunhofer IAIS. https://doi.org/10.24406/publica-fhg-300612.

[BL22] Blank, F. et al. 2022. Methodik zur Absicherung von KI im Fahrzeug. ATZ-Automobiltechnische Zeitschrift, 124(7).

[BMDV23] Bundesministerium für Digitales und Verkehr. 2023. G7 veröffentlichen Verhaltenskodex für Künstliche Intelligenz. (https://bmdv.bund.de/SharedDocs/DE/Pressemitteilungen/2023/109-wissing-g7-verhaltenskodex-ki.html, letzter Aufruf am 28.11.2023).

[BO21] Bommasani, R. et al. 2021. On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

[BOM23] Bommasani, R. et al. 2023. Do Foundation Model Providers Comply with the Draft EU AI Act? (https://crfm.stanford.edu/2023/06/15/eu-ai-act.html, letzter Aufruf am 05.12.2023).

[BO23] Boreiko, V. et al. 2023. Identifying Systematic Errors in Object Detectors with the SCROD Pipeline. IEEE/CVF International Conference on Computer Vision.

[BS23] Bundesamt für Sicherheit in der Informationstechnik. 2023. Große KI-Sprachmodelle - Chancen und Risiken für Industrie und Behörden. (https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/KI/Grosse_KI_Sprachmodelle.html, letzter Aufruf am 05.12.2023).

[CA21] Carlini, N. et al. 2021. Extracting training data from large language models. In 30th USENIX Security Symposium.

[CH23] Chen, L. et al. 2023. How is ChatGPT's behavior changing over time? arXiv preprint arXiv:2307.09009.

[CH23a] Chui, M. et. al. 2023. The economic potential of generative AI: The next productivity frontier. McKinsey. (https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier#/, letzter Aufruf am 05.12.2023).

[CL19] Clark, C. et al. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. arXiv preprint arXiv:1905.10044.

[CO21] Cobbe, K. et al. 2021. Training verifiers to solve math word problems. arXiv preprint arXiv:2110.14168.

[CR19] Cremers, A. et al. 2019. Vertrauenswürdiger Einsatz von Künstlicher Intelligenz. (https://www.ki.nrw/wp-content/uploads/2020/03/Whitepaper_KI-Zertifizierung.pdf, letzter Aufruf am 05.12.2023).

[DH21] Dhamala, J. et al. 2021. Bold: Dataset and metrics for measuring biases in open-ended language generation. In Proceedings of the 2021 ACM conference on Fairness, Accountability, and Transparancy.

[EU19] High-Level Expert Group. 2019. High-Level Expert Group on Artificial Intelligence. Ethics guidelines for trustworthy AI. Brüssel. (https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html, letzter Aufruf am 05.12.2023).

[EU23] European Parliament - Committee on the Internal Market and Consumer Protection, Committee on Civil Liberties, Justice and Home Affairs. Draft Compromise Amendments on the Draft Report Proposal for a regulation of the European Parliament and of the Council on harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts. 2023. (https://artificialintelligenceact.eu/wp-content/uploads/2023/06/AIA-%E2%80%93-IMCO-LIBE-Draft-Compromise-Amendments-14-June-2023.pdf, letzter Aufruf am 05.12.2023).

[FE20] Feldman, V. 2020. Does learning require memorization? A short tale about a long tail. ACM SIGACT Symposium on Theory of Computing.

[GA23] Gannamaneni, S., et al. 2023. Investigating CLIP Performance for Meta-Data Generation. In AD Datasets in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops.

[GE20] Gehman, S. et al. 2020. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. Conference on Emperical Methods in Natural Language Processing.

[GHC23] GitHub, Inc. 2023. GitHub Copilot. Your AI pair programmer. (https://github.com/features/copilot, letzter Aufruf am 28.11.2023).

[GL22] Glaese, A. et al. 2020. Improving alignment of dialogue agents via targeted human judgements. arXiv preprint arXiv:2209.14375.

[GÖ23] Görge, R. 2023. Generalized Measurement of Data Quality with Foundation Models. (interner Bericht).

[GO12] Gordon, A. et al. 2012. SemEval-2012 Task 7: Choice of Plausible Alternatives: An Evaluation of Commonsense Causal Reasoning. Conference on Lexical and Computational Semantics – Workshop on Semantic Evaluation.

[HAE23] Haedecke, E. G. et al. 2023. KI-Anwendungen systematisch prüfen und absichern. https://doi.org/10.24406/publica-1635.

[HE20] Hendrycks, D. et al. 2020. Measuring Massive Multitask Language Understanding. International Conference on Learning Representations.

[HE21] Hendrycks, D. et al. 2021. Measuring Mathematical Problem Solving With the MATH Dataset. NeurIPS 2021.

[HE23] Henderson, P. et al. 2023. Foundation models and fair use. arXiv preprint arXiv:2303.15715.

[JO17] Joshi, M. et al. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. Annual Meeting of the Association for Computational Linguistics.

[KO18] Kočiský, T. et al. 2018. The NarrativeQA Reading Comprehension Challenge. Transactions of the Association for Computational Linguistics, 6.

[KOM23] Europäische Kommission. 2023. Commission welcomes political agreement on Artificial Intelligence Act. Pressemitteilung, 9. Dezember 2023, Brüssel. (https://ec.europa.eu/commission/presscorner/detail/en/ip_23_6473, letzter Aufruf am 22.12.2023).

[LE21] Leslie, D. et al. 2021. Human rights, democracy, and the rule of law assurance framework for AI systems: A proposal. The Alan Turing Institute. https://doi.org/10.5281/zenodo.5981676.

[LE23] Große KI-Modelle für Deutschland. 2023. LEAM: AI, KI-Bundesverband. (https://leam.ai/wp-content/uploads/2023/01/LEAM-MBS_KIBV_webversion_mitAnhang_V2_2023.pdf, letzter Aufruf am 05.12.2023).

[LI21] Lin, S. et al. 2021. TruthfulQA: Measuring how models mimic human falsehoods. Annual Meeting of the Association for Computational Linguistics.

[LI22] Liang, P. et al. 2022. Holistic Evaluation of Language Models. arXiv preprint arXiv:2211.09110.

[MC23] Microsoft Corporation. 2023. Grounding LLMs. (https://techcommunity.microsoft.com/t5/fasttrack-for-azure/grounding-llms/ba-p/3843857, letzter Aufruf am 29.11.2023).

[ME23] Medeiros, L. 2023. Language Segment-Anything. (https://github.com/luca-medeiros/lang-segment-anything, letzter Aufruf am 15.11.2023).

[MI18] Mihaylov, T. et al. 2018. Can a suit of armor conduct electricity? A new dataset for open book question answering. Conference on Empirical Methods in Natural Language Processing.

[MI19] Mitchell, M. et al. 2019. Model cards for model reporting. ACM Conference on Fairness, Accountability and Transparancy.

[MO23] Mökander, J. et al. 2023. Auditing large language models: a three-layered approach. AI and Ethics.

[MRC20] Mattson, P. et al. 2020. MLPerf: An Industry Standard Benchmark Suite for Machine Learning Performance. IEEE Micro.

[MSC13] Mikolov, T. et al. 2013. Distributed representations of words and phrases and their compositionality. NeurIPS 2013.

[NA16] Nallapati, R. et al. 2016. Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond. Conference on Computational Natural Language Learning.

[NA18] Narayan, S. et al. 2018. Don't give me the details, just the summary! Topic-aware convolutional neural networks for extreme summarization. arXiv preprint arXiv:1808.08745.

[NA21] Nakano, R. et al. 2021. WebGPT: Browser-assisted question-answering with human feedback. arXiv preprint arXiv:2112.09332.

[NA23] National Cyber Security Centre. 2023. Guidelines for secure AI system development. (https://www.ncsc.gov.uk/

files/Guidelines-for-secure-AI-system-development.pdf, letzter Aufruf am 05.12.2023).

[NE21] Neel, A. et al. 2021. RAFT: A Real-World Few-Shot Text Classification Benchmark. NeurIPS 2021.

[NG22] Ngyen, N. et al. 2022. An Empirical Evaluation of GitHub Copilot's Code Suggestions. IEEE/ACM International Conference on Mining Software Repositories.

[NI23] National Institute of Standards and Technology (U.S. Department of Commerce). 2023. Artificial Intelligence Risk Management Framework (AI RMF 1.0). (https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf, letzter Aufruf am 14.02.2023).

[OA22] OpenAI, Inc. 2022. DALL·E-2. (https://openai.com/dall-e-2, letzter Aufruf am 29.11.2023).

[OB22] OpenAI, Inc. 2022. Reducing Bias and Improving Safety in DALL·E 2. (https://openai.com/blog/reducing-bias-and-improving-safety-in-dall-e-2/, letzter Aufruf am 05.12.2023).

[OE22] OECD. 2022. OECD Framework for the Classification of AI systems. OECD Digital Economy Papers, No. 323, OECD Publishing, Paris, https://doi.org/10.1787/cb6d9eca-en.

[PA21] Parrish, A. et al. 2021. BBQ: A hand-built bias benchmark for question answering. Annual Meeting of the Association for Computational Linguistics.

[PA23] Park, P. et al. 2023. AI Deception: A Survey of Examples, Risks, and Potential Solutions. https://arxiv.org/pdf/2308.14752.pdf.

[PAR23] Europäisches Parlament. 2023. Artificial Intelligence Act: deal on comprehensive rules for trustworthy AI. Pressemitteilung, 9. Dezember 2023. (https://www.europarl.europa.eu/news/en/press-room/20231206IPR15699/artificial-intelligence-act-deal-on-comprehensive-rules-for-trustworthy-ai, letzter Aufruf am 22.12.2023).

[PO20] Poretschkin, M. et al. 2020. Zur Systematischen Bewertung der Vertrauenswürdigkeit von KI-Systemen. doi.org/10.5771/9783748927990-175.

[PO21] Poretschkin, M. et al. 2021. KI-Prüfkatalog: Leitfaden zur Gestaltung vertrauenswürdige Intelligenz. Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS. (https://www.iais.fraunhofer.de/de/forschung/kuenstliche-intelligenz/ki-pruefkatalog.html, letzter Aufruf am 05.12.2023).

[RA19] Radford, A. et al. 2019. Language models are unsupervised multitask learners. OpenAI blog, https://github.com/openai/gpt-2.

[RA21] Radford, A. et al. 2021. Learning transferable visual models from natural language supervision. International conference on machine learning.

[RAJ18] Rajpurkar, P. et al. 2018. Know What You Don't Know: Unanswerable Questions for SQuAD. Annual Meeting of the Association for Computational Linguistics.

[RAT23] Europäischer Rat. 2023. Gesetz über künstliche Intelligenz: Rat und Parlament einigen sich über weltweit erste Regelung von KI. Pressemitteilung, 9. Dezember 2023. (https://www.consilium.europa.eu/de/press/press-releases/2023/12/09/artificial-intelligence-act-council-and-parliament-strike-a-deal-on-the-first-worldwide-rules-for-ai/, letzter Aufruf am 22.12.2023).

[RE23] Reuters. 2023. New York lawyers sanctioned for using fake ChatGPT cases in legal brief. (https://www.reuters.com/legal/new-york-lawyers-sanctioned-using-fake-chatgpt-cases-legal-brief-2023-06-22/, letzter Aufruf am 29.11.2023).

[SA22] Safe.trAIn. 2022. (https://safetrain-projekt.de/, letzter Aufruf am 14.11.2023).

[SA23] Center for AI Safety. 2023. Statement on AI Risk. (https://www.safe.ai/statement-on-ai-risk, letzter Aufruf am 05.12.2023).

[ST22] Steed, R. et al. 2022. Upstream Mitigation Is Not All You Need: Testing the Bias Transfer Hypothesis in Pre-Trained Language Models. Annual Meeting of the Association for Computational Linguistics.

[ST23] Stelzel, P. 2023. Die 8 besten KI-Prompt-Marktplätze zum Verkaufen von Midjourney- oder ChatGPT-Prompts. (https://philipp-stelzel.com/de/besten-prompt-marktplaetze/, letzter Aufruf am 17.11.2023).

[SU22] Suzgun, M. et al. 2022. Challenging BIG-Bench tasks and whether chain-of-thought can solve them. Annual Meeting of the Association for Computational Linguistics.

[UN22] Ung, M. et al. 2022. SaFeRDialogues: Taking Feedback Gracefully after Conversational Safety Failures. Annual Meeting of the Association for Computational Linguistics.

[WA18] Wang, A. et al. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. Conference on Empirical Methods in Natural Langague Processing Workshops.

[WA19] Wang, A. et al. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. NeurIPS 2019.

[WA22] Wang, X. et al. 2022. Self-consistency improves chain of thought reasoning in language models. ICLR 2022.

[WA23] Wang, B. et al. 2023. DecodingTrust: A Comprehensive Assessment of Trustworthiness of GPT Models. NeurIPS 2023.

[WE22] Wei, J. et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. NeurIPS 2022.

[WE22a] Wei, J. et al. 2022. Emergent abilities of large language models. Transactions on Machine Learning Research 08/2022.

[WH23] Whatplugin.AI. 2023. Number of ChatGPT Plugins. 2023. (https://www.whatplugin.ai/blog/chatgpt-plugins, letzter Aufruf am 29.11.2023).

[YA15] Yang, Y. et al. 2015. WikiQA: A challenge dataset for open-domain question answering. Conference on empirical methods in natural language processing.

[YA18] Yang, Z. et al. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. Conference on empirical methods in natural language processing.

[ZH19] Zhang T. et al. 2019. BERTScore: Evaluating Text Generation with BERT. International Conference on Learning Representations.

[ZLZ18] Zhu, Y. et al. 2018. Texygen: A Benchmarking Platform for Text Generation Models. ACM SIGIR Conference on Research & Development in Information Retrieval.

Impressum

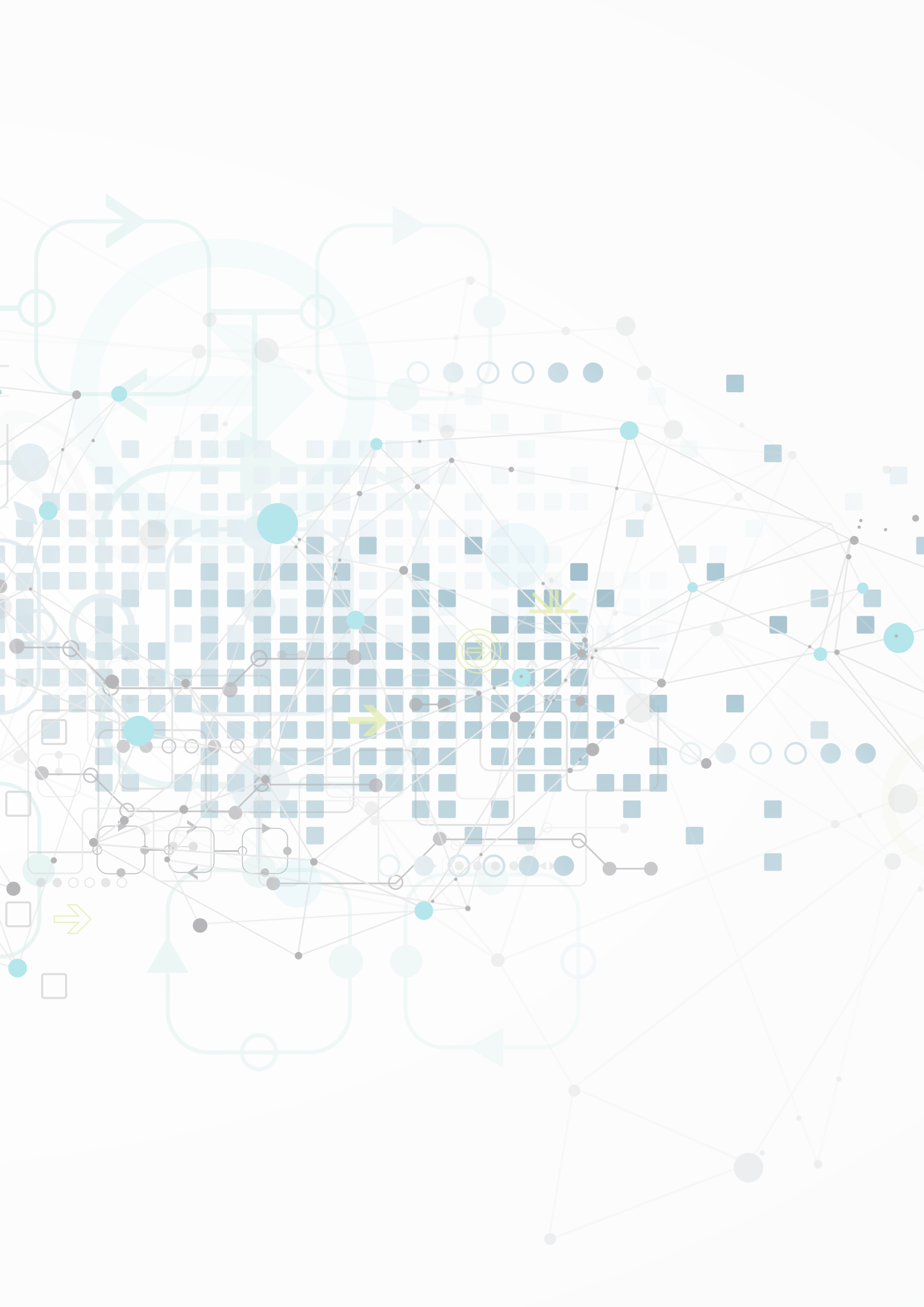
Herausgeber
Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS
Schloss Birlinghoven 1
53757 Sankt Augustin

Redaktion
Carmen Kern
Silke Loh

Grafik und Layout
Achim Kapusta
Eva Jüssen

Bildquellen
Titelbild: Alex – stock.adobe.com

Stand
Januar 2024



Kontakt

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

www.iais.fraunhofer.de

Ansprechpartner:
PD Dr. Michael Mock
michael.mock@iais.fraunhofer.de