Vertrauenswürdige Künstliche Intelligenz

10

Ute Schmid

Zusammenfassung

Für die sichere und sinnvolle Anwendung von KI-Systemen, insbesondere solchen, die auf komplexen, aus Daten gelernten Modellen basieren, müssen KI-Systeme vertrauenswürdig sein und Anwendende in die Lage versetzt werden, die Vertrauenswürdigkeit einschätzen zu können. Anforderungen an Vertrauenswürdige KI sind insbesondere Performanz und Robustheit, Transparenz und Erklärbarkeit, Diskriminierungsfreiheit sowie menschliche Kontrolle und Aufsicht. Um diese Anforderungen zu erfüllen, werden in der KI-Forschung Methoden entwickelt, die die Kernmethoden des maschinellen Lernens erweitern. Damit Nutzende ihr Vertrauen in KI-Systeme sinnvoll kalibrieren können, müssen die Schnittstellen zwischen KI-System und Mensch so gestaltet sein, dass die Ausgabe eines KI-Systems fundiert bewertet und gegebenenfalls korrigiert werden kann. Auf dieser Grundlage können partnerschaftliche KI-Systeme entwickelt werden, die Menschen dabei unterstützen, komplexe Probleme effizient und angemessen zu lösen.

Die Entwicklung von hochperformanten Methoden und Architekturen im Bereich maschinelles Lernen eröffnet Einsatzmöglichkeiten von KI-Methoden in immer mehr Anwendungsbereichen. Während in einigen Bereichen voll autonome KI-Systeme entwickelt werden, etwa für autonomes Fahren, werden in der überwiegenden Zahl von Anwendungsfeldern Systeme zum Einsatz kommen, bei denen die letztendliche Entscheidung beim Menschen liegt. Dies gilt für Klassifikationssysteme ebenso wie für generative KI-

U. Schmid (✉)
Lehrstuhl für Kognitive Systeme, Universität Bamberg, Bamberg, Deutschland
E-Mail: ute.schmid@uni-bamberg.de

F. Schmiedchen et al. (Hrsg.), *Künstliche Intelligenz und Wir*,
https://doi.org/10.1007/978-3-662-71567-3_10

Systeme. Beispielsweise kann bildbasierte medizinische Diagnose durch Modelle unterstützt werden, die mit CNNs (*convolutional neural networks*, Krizhevsky et al., 2012a, b) auf Bilddaten trainiert wurden. Die diagnostische Entscheidung sowie die Auswahl einer geeigneten Therapie sollten allerdings beim medizinischen Fachpersonal verbleiben (Bruckert et al., 2020). Ein Text, der mit einem auf einem großen Sprachmodell (*large language model*, LLM, Shanahan, 2024) basierenden generativen KI-System erzeugt wurde, sollte von Menschen überprüft und gegebenenfalls korrigiert werden (Gao et al., 2024).

Während bei Standardsoftware zumindest theoretisch gewährleistet werden kann, dass die zugrunde liegenden Programme korrekt und vollständig sind, ist dies bei KI-Systemen nicht der Fall. Ein Programm ist korrekt, wenn es für alle Eingaben garantiert die richtige Ausgabe liefert, und es ist vollständig, wenn dies für alle möglichen Eingaben der Fall ist (Zowghi & Gervasi, 2002). KI-Methoden ermöglichen es, Probleme mit Computern zu bearbeiten, die durch Standard-Algorithmen nicht lösbar sind. Insbesondere gilt das für die folgenden drei Fälle (s. Schmid, 2024):

1. Ein Problem ist so komplex, dass seine Lösung nicht effizient berechenbar ist. Dies gilt beispielsweise für viele Spiele, etwa Schach oder Go, sowie für das Finden optimaler Wege. In diesem Fall werden heuristische Methoden genutzt, die erlauben abzuschätzen, welche Lösungswege vielversprechend sind und welche nicht. Allerdings kann dann nicht garantiert werden, dass das Programm die beste Lösung (oder sogar überhaupt eine Lösung) findet.
2. Ein Problembereich basiert auf komplexem Domänenwissen und allgemeinem Wissen (*common sense*) sowie der Notwendigkeit, Schlussfolgerungen aus diesem Wissen zu ziehen. Hier sind Datenstrukturen und Algorithmen notwendig, die über Standardmethoden der Informatik hinausgehen. Dies ist das Einsatzgebiet wissensbasierter Systeme.
3. Schließlich gibt es Probleme, die nicht vollständig oder gar nicht explizit beschrieben werden können. Dies ist für Problembereiche der Fall, bei denen Menschen implizites Wissen haben. Dies sind automatisierte Entscheidungsroutinen und Strategien sowie insbesondere perzeptuelles Wissen. Beispielsweise ist es unmöglich, vollständig zu beschreiben, welche Regeln wir verwenden, um auf einem Bild zu erkennen, ob eine Katze darauf abgebildet ist oder ob es sich bei einer Hautveränderung um Hautkrebs handelt. Hier kommen Methoden des maschinellen Lernens zur Anwendung, mit denen aus Beispieldaten Modelle generalisiert werden. Von Menschen erstellte Programme, die für gegebene Eingaben die passenden Ausgaben berechnen, werden durch meist intransparente (*black box*) Modelle ersetzt.

Mittels dieser mächtigen Familien von KI-Methoden ist es möglich, viele Probleme mit Computern zu lösen, die zuvor nur von Menschen lösbar waren. Genau dies entspricht der klassischen Definition von Künstlicher Intelligenz (Rich, 1983). Künstliche Intelligenz war lange ein Teilgebiet der Informatik, bei dem der Fokus vor allem auf Grundlagenforschung lag. Für Standardsoftware existieren dagegen eine lange Tradition in der Entwicklung von Anwendungssystemen und entsprechende Methoden zur Prüfung von Software-

qualität (Balzert, 1998). Mit der zunehmenden Anwendungsrelevanz von KI entstand auch der Bedarf, die Qualität von KI-Systemen bewerten zu können. Ein zentraler Beitrag hierfür sind die Anforderungen an Vertrauenswürdige KI, die eine Gruppe von Expertinnen und Experten im Auftrag der europäischen Kommission entwickelt hat (HEG-KI, 2019). Im Folgenden werden diese Anforderungen vorgestellt und der Zusammenhang von Vertrauenswürdigkeit eines Systems und menschlichem Vertrauen in ein System diskutiert. Nachfolgenden werden vier der Anforderungen vertieft behandelt (siehe auch Schmid, 2022, 2024). Für diese Anforderungen wurden in den letzten Jahren neue KI-Methoden entwickelt, die die jeweiligen Kerntechnologien erweitern und ergänzen. Die methodischen Entwicklungen haben sich zunächst auf die Vertrauenswürdigkeit von Klassifikationssystemen fokussiert, insbesondere solche, die auf komplexen neuronalen Netzen (*deep learning*) basieren. Aktuell werden entsprechende Methoden für generative KI entwickelt. Hier steht die Forschung jedoch noch am Anfang. Auch wenn die Anforderungen an Vertrauenswürdigkeit vor allem auf KI-Systeme bezogen sind, die auf maschinellem Lernen basieren, können diese auch auf KI-Systeme, die auf wissensbasierten Methoden basieren, angewendet werden.

10.1 Vertrauenswürdigkeit von und Vertrauen in KI-Systeme

Mit der wachsenden Zahl an Anwendungen von KI-Methoden in immer mehr Lebens- und Arbeitsbereichen ergibt sich die Notwendigkeit, prüfbare Kriterien zu entwickeln, die es erlauben, die Vertrauenswürdigkeit von KI-Systemen zu bewerten. Entsprechend haben verschiedene Institutionen, darunter die International Organization for Standardization (ISO), das U.S. Government Accountability Office (GAO) und die Europäische Union Programme aufgelegt, um entsprechende Leitlinien zu entwickeln (Kaur et al., 2022). Die von den verschiedenen Institutionen vorgelegten Kriterien zeigen hohe Übereinstimmung. Viel beachtet sind die von der Europäischen Kommission vorgelegten sieben Anforderungen (HEG-KI, 2019), die im Folgenden vorgestellt werden. Leitgedanke war hier die Gewährleistung eines angemessenen ethischen und rechtlichen Rahmens zur Stärkung der europäischen Werte mit der Vision, dass KI-Systeme entstehen, die dazu beitragen können, die Ziele für nachhaltige Entwicklung der Vereinten Nationen zu erreichen, beispielsweise bei der Bekämpfung des Klimawandels, beim rationalen Umgang mit natürlichen Ressourcen, bei der Gesundheitsförderung und bei der Geschlechtergerechtigkeit. In den Leitlinien wurde konstatiert, dass den vielfältigen Chancen, die sich durch die Nutzung von KI-Systemen ergeben, Risiken gegenüberstehen, die angemessen und verhältnismäßig behandelt werden sollten. Es soll gewährleistet werden, dass den sozio-technischen Umgebungen, in die KI-Systeme eingebettet sind, vertraut werden kann, und erreicht werden, dass KI-Unternehmen durch die Vertrauenswürdigkeit ihrer Produkte und Dienstleistungen einen Wettbewerbsvorteil erlangen.

Die sieben Anforderungen, die die HEG-KI formuliert hat, sind in Tab. 10.1 zusammengefasst. Dabei werden sowohl technische als auch nichttechnische Anforderungen

Tab. 10.1 Anforderungen an Vertrauenswürdige KI-Systeme. (HEG-KI, 2019)

	Anforderung	Bereich	KI-Methoden
1	Vorrang menschlichen Handelns und menschliche Aufsicht	KI und Mensch-Computer-Interaktion	Erklärbarkeit und Interaktivität
2	Technische Robustheit und Sicherheit	KI und Software Engineering	Performanzevaluation und hybride KI
3	Schutz der Privatsphäre und Datenqualitätsmanagement	Informatik und Recht	
4	Transparenz	KI, Kognitionswissenschaft	Erklärbarkeit
5	Vielfalt, Nichtdiskriminierung und Fairness	KI	Bias-Vermeidung und -Reduktion
6	Gesellschaftliches und ökologisches Wohlergehen	Sozio-technische Einbettung	
7	Rechenschaftspflicht	Rechtswissenschaft	

formuliert. Als erste Anforderung wird der Vorrang menschlichen Handelns und menschliche Aufsicht genannt. KI-Systeme sollten Menschen dabei unterstützen, fundierte Entscheidungen treffen zu können, die im Einklang mit ihren eigenen Zielen stehen. Unfaire Formen der Manipulation, die die menschliche Autonomie gefährden, sollen entsprechend vermieden werden. Menschliche Aufsicht kann durch die Möglichkeit zur Überprüfung und Kontrolle der Prozesse und Ausgaben eines KI-Systems oder durch interaktive Einbindung des Menschen (*human-in-the-loop*) gewährleistet werden. Die Kontrollierbarkeit von Prozessen und Ausgaben ist eng mit der Anforderung an Transparenz, insbesondere der Erklärbarkeit, verbunden. Die zweite Anforderung adressiert die technische Robustheit und Sicherheit von KI-Systemen. Diese Anforderung entspricht einer Übertragung von Prinzipien guter Software auf KI-Systeme. Hier geht es um die Vermeidung von Sicherheitslücken sowie die Präzision und Zuverlässigkeit von aus Daten gelernten Modellen. Die dritte Anforderung behandelt den Schutz der Privatsphäre und das Datenqualitätsmanagement. Hier geht es um das Einhalten von Datenschutzvorgaben sowie die Qualität der Datensätze, mit denen Modelle trainiert werden. Diese Anforderung hat auch Bezüge zur fünften Anforderung der Nichtdiskriminierung und der Vermeidung von unerwünschten Verzerrungen in den Trainingsdaten. Als vierte Anforderung wird Transparenz genannt. Transparenz umfasst die Offenlegung von Trainingsdaten und der genutzten Algorithmen, den klaren Ausweis, wenn eine Kommunikation mit einem KI-System und nicht mit einem Menschen stattfindet, sowie die Erklärbarkeit. Erklärbarkeit von KI-Systemen meint, dass die von einem KI-System getroffenen Entscheidungen von Menschen verstanden und rückverfolgt werden können. Zur Transparenz sollte zudem die Information gehören, wie viele Data Worker zu welchen Löhnen beschäftigt wurden, um die Daten, mit denen Modelle trainiert wurden, zu annotieren und geeignet aufzubereiten, und wie viel Zeit an menschlicher Arbeit in das Training der Modelle geflossen ist – etwa bei den zeitintensiven Arbeiten für Dialogtraining für große Sprachmodelle. Die fünfte Anforderung bezieht sich auf Vielfalt, Nichtdiskriminierung und Fairness. Hier geht es um die Vermeidung der Benachteiligung bestimmter Personengruppen, beispielsweise weil diese

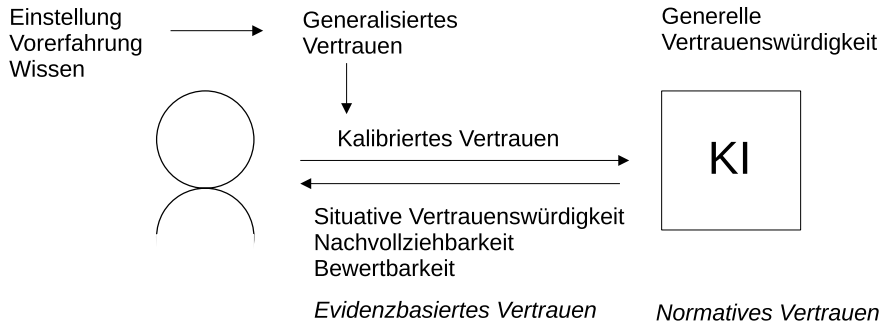

Abb. 10.1 Die Beziehung von Vertrauenswürdigkeit von und Vertrauen in KI-Systeme

in den Trainingsdaten unterrepräsentiert sind. Zudem sind die für Software generell gültigen Anforderungen an Barrierefreiheit sowie die Beteiligung aller relevanten Interessensgruppen beim Entwurf der Systeme hier adressiert. Als sechste Anforderung wird gesellschaftliches und ökologisches Wohlergehen genannt. KI-Systeme sollen möglichst nachhaltig und umweltfreundlich sein, nicht zu negativen sozialen Auswirkungen beitragen und keine unerwünschten Effekte auf demokratische Meinungsbildungsprozesse haben. Schließlich ist die siebte Anforderung die Rechenschaftspflicht (*accountability*), insbesondere die Nachprüfbarkeit und Berichterstattung, um bei schädlichen Auswirkungen die Verantwortlichkeit klären zu können.

Die Realisierung dieser sieben Anforderungen an Vertrauenswürdigkeit verlangt sowohl nichttechnische Methoden, insbesondere regulatorische Maßnahmen, als auch technische Methoden. Die technischen Methoden können teilweise direkt aus dem Bereich der Standardsoftware übernommen werden, etwa Anforderungen an die Datensicherheit. Teilweise müssen neue Methoden, darunter auch neue KI-Methoden, entwickelt werden, um den Anforderungen zu begegnen. In Tab. 10.1 wird eine grobe Zuordnung von einschlägigen Bereichen zu Anforderungen vorgeschlagen. Für vier der sieben Anforderungen wurden in den letzten Jahren KI-Methoden entwickelt, die in den folgenden Unterkapiteln beschrieben werden.

Die Vertrauenswürdigkeit eines KI-Systems sollte die Grundlage dafür liefern, ob Menschen einem KI-System vertrauen oder vertrauen sollten. Vertrauenswürdigkeit ist also eine Eigenschaft des Systems, Vertrauen eine Zuschreibung von Vertrauenswürdigkeit an ein System (s. Abb. 10.1). Eine dazu analoge Definition wird aus sozialpsychologischer Perspektive für interpersonelles Vertrauen gegeben (Robbins, 2016). Ähnliche Mechanismen werden für Vertrauen in Institutionen und Organisationen (Creed et al., 1996) oder Berufsgruppen (Hall et al., 2001) identifiziert. Das sozialpsychologische Vertrauenskonzept wurde auf die Messung von Vertrauen in Technologien (Mcknight et al., 2011; Sheridan, 2019) sowie in den letzten Jahren auch auf Vertrauen in KI-Systeme (Holliday et al., 2016; Glikson & Woolley, 2020; Kaplan et al., 2023) übertragen. Faktoren, die das Vertrauen in KI-Systeme beeinflussen, lassen sich nach Kaplan et al. (2023) in vier Gruppen einteilen, zu denen unter anderem folgende Einflussgrößen gehören:

- **Art des KI-Systems:** Algorithmus, Chatbot, Roboter, autonomes Fahrzeug
- **Menschbezogene Aspekte:** Vorerfahrung, Bildung, Einstellung zu KI, generelle Tendenz zu vertrauen
- **Eigenschaften des KI-Systems:**
  - performanzbasiert: Verlässlichkeit, Vorhersagbarkeit, Zuverlässigkeit
  - merkmalsbasiert: Grad der Anthropomorphisierung, Grad der Autonomie, Transparenz
- **Kontext:** Kritikalität der Aufgabe, Komplexität der Aufgabe

Skalen zur Messung des Vertrauens in KI-Systeme, wie die *General Attitudes towards Artificial Intelligence Scale* (GAAIS, Schepman & Rodway, 2023), erfassen eine generalisierte Vertrauenszuweisung in KI-Systeme. Insbesondere im Kontext der Forschung zu Erklärbarer KI liegt der Fokus auf der Gestaltung von Mensch-KI-Schnittstellen, die ein situativ adaptives, kalibriertes Vertrauen in eine spezifische Ausgabe eines KI-Systems im Kontext einer speziellen Aufgabe adressieren (Thaler & Schmid, 2021; Tomsett et al., 2020; Zhang et al., 2020). Die Beziehung zwischen der Charakterisierung eines speziellen KI-Systems in Bezug auf dessen Erfüllung der sieben Anforderungen an Vertrauenswürdigkeit und der spezifischen Erfahrung von Nutzenden mit KI-Systemen allgemein sowie einem speziellen KI-System im Aufgabenkontext kann als normatives Vertrauen einerseits und evidenzbasiertes Vertrauen andererseits klassifiziert werden. Evidenzbasiertes Vertrauen umfasst dabei situativ kalibriertes Vertrauen sowie vorangegangene Erfahrung mit diesem und anderen KI-Systemen.[1]

10.2 Performanz und Robustheit

Im Folgenden werden die verschiedenen Anforderungen an die Vertrauenswürdigkeit von KI-Systemen mit Fokus auf das Lernen von Klassifikatoren diskutiert. Klassifikationslernen umfasst verschiedene überwachte Ansätze des maschinellen Lernens (*supervised learning*), bei denen mittels einer Menge von annotierten Trainingsdaten ein Modell aufgebaut wird (z. B. Lindholm et al., 2022). Annotation meint, dass zu jedem Beispiel der Trainingsmenge die korrekte Klasse mitgegeben wird. Das Modell lernt eine Abbildung von Eingabedaten auf die korrekte Klasse und nutzt dabei die vorgegebene Klasseninformation zur Modellanpassung. Die Zuweisung der korrekten Klassen (*labeling*) wird in den meisten Fällen durch Menschen erledigt. Häufig wird übersehen, dass hinter vielen KI-Systemen ein enormer Aufwand an menschlicher Arbeit steckt. Die großen KI-Unternehmen beschäftigen sehr viele solche Data Worker unter oft prekären Bedingungen (Williams et al., 2022). Für Klassifikatoren in hochspezialisierten Bereichen, etwa der me-

[1] Die Unterscheidung von normativem und evidenzbasiertem Vertrauen stammt von Dirk Heckmann (bidt) im Kontext einer Diskussion im Rahmen des bidt-Forschungsschwerpunkts Mensch und generative Künstliche Intelligenz: Trust in Co-Creation, 11.10.2024.

dizinischen Diagnostik, müssen zum Labeling Expertinnen und Experten herangezogen werden. In Bereichen, wo das Labeling nicht eindeutig ist, also keine *ground truth* existiert, annotieren wenn möglich mehrere Personen die gleichen Daten, und es werden ähnliche Methoden genutzt wie bei der statistischen Analyse qualitativer Daten (Chew et al., 2019). Da die Performanz der gelernten Modelle maßgeblich von der Qualität der genutzten Trainingsdaten abhängt, sind qualitativ hochwertige Datensätze entsprechend wertvoll.

Im Kontext des maschinellen Lernens hat sich für die Performanzbeurteilung von gelernten Modellen im Bereich der Klassifikation eine allgemein akzeptierte Methodik entwickelt (Lindholm et al., 2022, Kap. 4). Ein Teil der vorhandenen Daten wird nicht zum Training des Modells genutzt, sondern als Testdatenmenge zurückbehalten. Nachdem ein Modell trainiert wurde, wird es auf die Testdaten angewendet, für die aber ebenfalls bereits die gewünschte Klassenausgabe vorgegeben ist. Nun kann für die Testdaten beobachtet werden, wie oft das Modell eine korrekte oder fehlerhafte Klasse liefert. Mit dieser Information wird die prädiktive Performanz des Modells abgeschätzt, also wie gut das Modell für neue, noch nicht gesehene Eingaben funktionieren wird. Üblicherweise werden hier die Präzision und die Sensitivität (*recall*) betrachtet. Präzision erfasst den Anteil an korrekt klassifizierten Eingaben relativ zu allen Eingaben, die einer bestimmten Klasse zugeordnet wurden, also der Anzahl korrekt und falsch positiver Klassifikationen. Sensitivität erfasst den Anteil korrekt klassifizierter Eingaben relativ zur Menge aller Eingaben, die mit dieser Klasse annotiert sind. Beide Maße werden häufig zu einem Gesamtscore (*F1 score*) verrechnet. Ist die Performanz eines Modells auf den Trainingsdaten höher als auf den Testdaten, spricht man von Überanpassung (*overfitting;* Ditterich, 1995; Rice et al., 2020). Das Modell nutzt dann irrelevante Merkmale, die spezifisch für die Trainingsdaten sind und mit der vorherzusagenden Klasse korrelieren (*spurious correlation*), und kann dann nicht mehr gut auf neue Daten generalisieren. Ist die Datengrundlage nicht repräsentativ für die Verteilung von Daten, bestehen diese irrelevanten Korrelationen allerdings auch in den Testdaten. Man trainiert damit sogenannte Kluge-Hans-Modelle,[2] die scheinbar hochperformant sind, aber in Wirklichkeit kein Modell zur Vorhersage der Zielklassen gelernt haben. Erklärbare KI bietet Möglichkeiten, solche Kluge-Hans-Modelle zu identifizieren (Lapuschkin et al., 2019).

Für die Bewertung der Performanz von Ansätzen der generativen KI gibt es aktuell noch kein etabliertes methodisches Vorgehen. Entsprechend dominieren eher anekdotische Erfahrungsberichte. Systematische empirische Evaluationen basieren entweder auf dem Vergleich der Übereinstimmung generierter Inhalte mit vorgegebenen Inhalten (Mizrahi et al., 2024) oder der Bewertung durch Menschen. Beispielsweise haben Herbold et al. (2023) von ChatGPT generierte Aufsätze und von Schülerinnen und Schülern ge-

[2] Der Begriff „Kluger Hans" bezieht sich auf ein Pferd, das angeblich zählen und rechnen konnte und zu Beginn des 20. Jahrhunderts für Aufmerksamkeit sorgte. Es stellte sich schließlich heraus, dass es auf wohl unbeabsichtigte Signale seines Besitzers, etwa subtile Änderungen der Körperhaltung, reagierte.

schriebene Aufsätze durch Lehrkräfte beurteilen lassen. Dabei war nicht gekennzeichnet, ob der Aufsatz von Mensch oder Maschine stammt. Hier zeigte sich, dass Aufsätze von ChatGPT-3.5 im Schnitt schlechter, Aufsätze von ChatGPT-4 aber besser als die von Schülerinnen und Schülern bewertet wurden.

Die Robustheit eines Klassifikationsmodells betrifft dessen Performanz für neue Eingaben (Freiesleben & Grote, 2023), insbesondere wenn sich die Verteilung von Daten ändert (*concept drift*), wenn Eingaben verrauscht oder manipuliert werden (*adversarial examples*) und wenn Eingaben erfolgen, die außerhalb des Bereichs liegen, mit denen ein Modell trainiert wurde (*out-of-distribution error*). Die Robustheit eines Modells betrifft also dessen Generalisierungsfähigkeit über die bereits gesehenen Daten hinaus. Fehlerhafte Ausgaben eines Modells können zu einem Verlust an Vertrauen führen, besonders dann, wenn der Fehler für Menschen offensichtlich ist. Beispielsweise ist für Menschen ein Stoppschild, auf das jemand einen Aufkleber angebracht hat, immer noch als solches zu erkennen; bei einem gelernten Modell kann dies zu einer Fehlklassifikation führen. Wurde ein Modell mit Tierbildern trainiert und ist entsprechend auch nur für die Klassifikation von Tieren vorgesehen, würde es auf die Eingabe eines anderen Bildes, beispielsweise eines Kühlschrankes, aufgrund der dominanten Farbe Weiß mit der Ausgabe einer Klasse aus dem Trainingsbereich, beispielsweise Eisbär, reagieren. Menschen würden bei einem unbekannten Beispiel dagegen erkennen, dass sie dazu noch kein Wissen haben.

Auf der einen Seite zeigen aus Daten gelernte Modelle inzwischen höhere Performanz als die meisten Menschen – etwa bei der Hautkrebs-Erkennung (Brinker et al., 2019). Auf der anderen Seite ist gerade für Alltagsbereiche die menschliche Generalisierungsfähigkeit deutlich robuster, flexibler und datensparsamer als maschinelles Lernen (Ilievski et al., 2024). Ein möglicher Zugang, um bessere Generalisierungsfähigkeit und mehr Robustheit zu erreichen, wird in der Kombination aus maschinellem Lernen und wissensbasierten KI-Methoden gesehen. Forschung zur Kombination von Wissen und Lernen wird als hybride KI oder als neurosymbolische KI (Sarker et al., 2022; Marra et al., 2024) bezeichnet (s. Abb. 10.2). Die Einbeziehung von Wissen kann maschinelles Lernen datensparsamer machen (siehe Schmid, 2024). Ein rein datengetriebenes Modell ist gezwungen, bestimmte Konzepte wieder und wieder aus Daten zu induzieren. Menschliches Lernen zeichnet sich dagegen dadurch aus, dass bereits vorhandenes Wissen im Lernprozess genutzt wird. So muss beispielsweise nicht immer wieder neu gelernt werden, dass Säugetiere Augen haben. Zudem reichern Menschen neue Beobachtungen häufig durch Schlussfolgerungen an. Sehe ich ein mir unbekanntes Tier, das Augen hat, so schließe ich daraus, dass es sehen kann. Umgekehrt können die hochperformanten Architekturen des tiefen Lernens dazu beitragen, wissensbasierte Ansätze flexibler zu machen. Fest vorgegebene Wissensbasen können durch Lernen erweitert und adaptiert werden. Explizit repräsentiertes Wissen kann mit implizitem Wissen kombiniert werden. Ein Beispiel hierfür ist DeepProblog (Manhaeve et al., 2021). Hier kann das Erkennen visueller Objekte (zum Beispiel handgeschriebene Ziffern) mit dem Lernen kognitiver Regeln (zum Beispiel für arithmetische Operationen) kombiniert werden. Einen ähnlichen Ansatz verfolgen Rabold et al. (2020) mit der Kombination von Bildklassifikation und dem Lernen relationaler Regeln

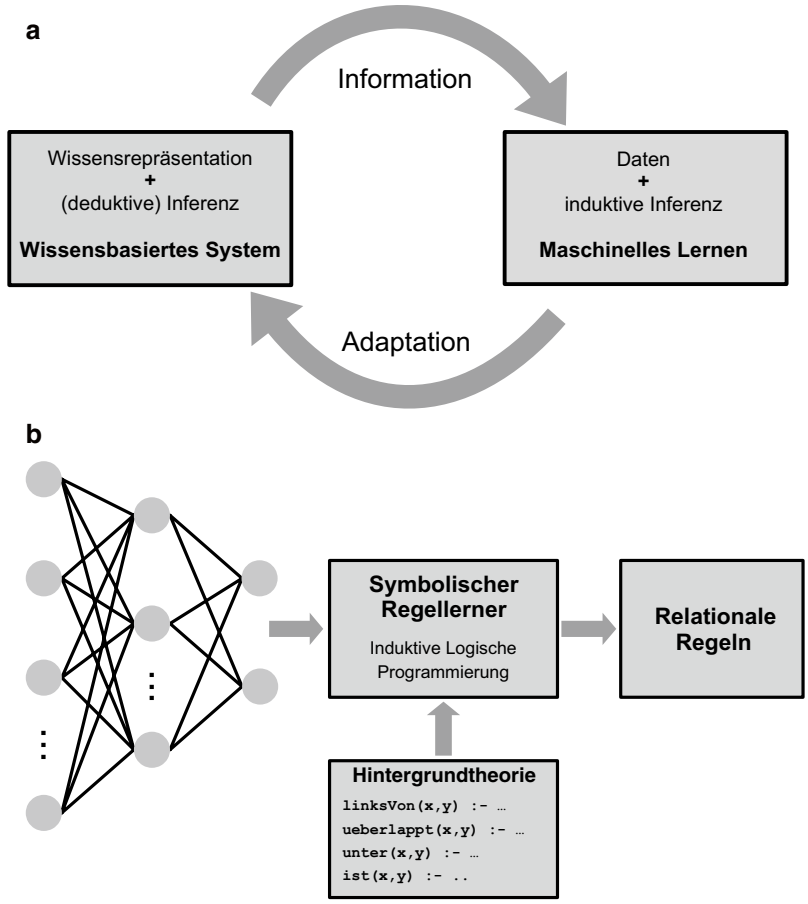

Abb. 10.2 Kombination von wissensbasierten Systemen und maschinellem Lernen zu einer hybriden Architektur (**a**) und Beispielarchitektur für ein neurosymbolisches KI-System, bei dem implizites Lernen mit neuronalen Netzen mit dem Lernen symbolischer, relationaler Regeln kombiniert wird (**b**)

mit induktiver logischer Programmierung (ILP). Beispielsweise können verschiedene Gebäudearten unterschieden werden, indem zunächst Fenster erkannt und darauf aufbauend Regeln gelernt werden, die die räumlichen Beziehungen zwischen Fenstern nutzen, um etwa einen Turm von einem Bungalow zu unterscheiden. Die Kombination von Bildklassifikation auf Basis neuronaler Netze mit dem Lernen relationaler Regeln ist gleichzeitig ein Beitrag zur Erklärbaren KI, da dadurch ermöglicht wird, dass komplexe Entscheidungsregeln symbolisch repräsentiert und sprachlich vermittelt werden können (Schmid, 2021).

Im Bereich der generativen KI werden ebenfalls Ansätze entwickelt, bei denen wissensbasierte Methoden mit neuronalen Netzen kombiniert werden. Insbesondere wird *retrieval augmented generation* (RAG, Gao et al., 2023) genutzt, um die Generierung von Inhalten zu augmentieren. Damit sollen fehlerhafte Ausgaben („Halluzinieren“) möglichst ver-

mieden werden. Zudem wird erreicht, dass die generierten Inhalte gezielt spezifische Informationen nutzen. Beispielsweise werden hier Wissensgraphen genutzt (Pan et al., 2024; Schramm et al., 2023).

10.3 Diskriminierungsfreiheit

Die Anforderung, dass ein KI-System möglichst diskriminierungsfrei ist (*fair AI,* Ruggieri et al., 2023), soll gewährleisten, dass bestimmte Personengruppen bezogen auf beispielsweise Geschlecht, Ethnie oder Alter nicht benachteiligt werden. Bei aus Daten gelernten Modellen hängt es maßgeblich von den Daten ab, mit denen ein Modell trainiert wurde, ob es diskriminierungsfrei ist oder nicht. Dies ist besonders dann der Fall, wenn eine bestimmte Personengruppe bezogen auf ein bestimmtes Kriterium unterrepräsentiert ist. Beispielsweise wies das KI-System, das Amazon 2015 zur Identifikation von geeigneten Bewerbungen auf Stellen genutzt hat, einen unfairen Gender-Bias auf, da Bewerbungen von Frauen auf Stellenangebote im Bereich Software Engineering systematisch nicht berücksichtigt wurden (Stahl et al., 2022). Die Ursache war, dass in den genutzten Trainingsdaten kaum Fälle von Softwareentwicklerinnen vorhanden waren. Die genutzten historischen Daten waren also verzerrt – es lag eine Stichprobenverzerrung (*sampling bias*) vor.

Im Bereich maschinelles Lernen existieren allerdings bereits seit Langem Methoden, um mit dem Problem von Gruppen umzugehen, die für bestimmte Zielgrößen unterrepräsentiert sind (Maloof, 2003), die im Fall von Amazon schlicht ignoriert wurden. Beispielsweise existieren verschiedene *resampling* und *reweighting* Methoden, um die Unterrepräsentation verschiedener Gruppen auszugleichen (Kamiran & Calders, 2012). Dies setzt allerdings voraus, dass vor dem Training der Modelle sensible Merkmale, die zu Diskriminierung führen können, identifiziert werden. Dies gilt auch für eine weitere Strategie, nämlich die Entfernung sensibler Merkmale aus den Daten, beispielsweise Geschlecht, weil mit diesem Merkmal hochkorrelierte Merkmale (bei Geschlecht etwa Merkmale wie Größe und Gewicht) ebenfalls zu unfairen Modellen führen können (Pahl et al., 2022). Teilweise sind unfaire Verzerrungen auch schwer in Gänze auszuräumen. Dies war zum Beispiel bei der Google Photo-App der Fall, bei der dunkelhäutige Menschen als Gorillas klassifiziert wurden (Lee, 2018). Das Problem wurde zeitweise dadurch umgangen, dass die Klasse Gorilla generell herausgefiltert wurde. Teilen sich Bilder in einem Datensatz, die zu verschiedenen Klassen gehören, viele Merkmale, lassen sich solche Fehlklassifikationen nicht generell vermeiden.

Auch bei generativer KI, beispielsweise bei Bildgeneratoren und bei maschineller Übersetzung, führt die Datengrundlage, mit denen die Modelle trainiert werden, zu unerwünschten Verzerrungen. Beispielsweise sind Frauen in generierten Bildern häufig jung und stereotyp attraktiv dargestellt, da in der Datengrundlage vermutlich viele editierte Bilder aus sozialen Medien enthalten sind. Um ethnische Verzerrungen möglichst zu vermeiden, hat Google bei Gemini explizit Algorithmen verwendet, die entsprechenden Verzerrungen entgegenwirken sollen. Dadurch entstehen allerdings völlig unplausible Dar-

stellungen, wie dunkelhäutige Menschen und Frauen in der Uniform deutscher Soldaten im zweiten Weltkrieg (Kleinman, 2024). Bei der maschinellen Übersetzung führen Korrelationen von Geschlecht und Berufsgruppen nicht nur zu unerwünschten Verzerrungen, sondern sogar zu fehlerhaften Texten. Beispielsweise wird der englischsprachige Satz *The cleaner hates the developer because she alwaysleaves the room dirty.* ins Deutsche übersetzt mit *Die Reinigungskraft hasst den Entwickler, weil sie das Zimmer immer schmutzig hinterlässt.* (Troles & Schmid, 2021).

Die Philosophin Shannon Vallor (2024) zeigt auf, dass unfaire Modelle uns einen Spiegel vorhalten und historische wie bestehende Ungerechtigkeiten aufzeigen. Allerdings sollte bedacht werden, dass es kaum Bereiche gibt, bei denen objektiv und allgemeingültig festgelegt werden kann, was fair und gerecht ist. Gerechtigkeit und Fairness sind abhängig vom kulturellen Kontext – sei es bezogen auf Länder, Unternehmen oder Institutionen. Entsprechend ist die Transparenz von KI-Systemen wichtig, damit menschliche Entscheiderinnen und Entscheider nachvollziehen können, auf Grundlage welcher Information ein System seine Ausgabe generiert – sei es bezogen auf die Vergabe eines Kredits oder die Entscheidung über eine medizinische Behandlung. Ein entsprechendes System, das Fairness im Kontext von erklärbarem interaktivem Lernen adressiert, wurde beispielsweise von Heidrich et al. (2023) vorgeschlagen.

10.4 Transparenz und Erklärbarkeit

Transparenz als Anforderung für die Vertrauenswürdigkeit von KI-Systemen wird insbesondere durch Methoden der Erklärbaren Künstlichen Intelligenz (*eXplainable AI*, XAI) realisiert. Der Beginn des Forschungsgebiets XAI kann auf das Jahr 2016 festgelegt werden, als David Gunning den Begriff im Rahmen eines Vortrags auf der *International Joint Conference on AI* (IJCAI) einführte (Gunning & Aha, 2019). Gunning argumentierte, dass es wünschenswert wäre, dass intransparente Modelle, besonders Bildklassifikatoren, zusätzlich zur Ausgabe einer Klassifikation auch eine Erklärung liefern würden, warum das Modell zu einer speziellen Klassenentscheidung kam. Am Beispiel der Klassifikation einer Katze illustrierte er eine mögliche Erklärung. Diese Erklärung war multimodal und angelehnt an menschliche Erklärungen: „Ich erkenne eine Katze, weil dort Schnurrhaare und Krallen zu sehen sind und die Ohren ähnlich zu folgenden prototypischen Bildern von Ohren sind, die zu Katzen gehören." Die Erklärung, für die es zu dieser Zeit keine algorithmische Methode gab, enthielt also einerseits sprachlich formulierbare Konzepte und andererseits prototypische Bilder von Katzenohren. In der Folge entstand eine wachsende Menge an XAI-Methoden. Allerdings lag der Fokus zunächst auf Methoden der Merkmalsrelevanz (*feature attribution methods*). Diese Methoden heben bei Klassifikatoren für Bilddaten beispielsweise in Form einer *heatmap* hervor, welche Pixel im Eingabebild vor allem vom Modell genutzt wurden, um die Ausgabeklasse zu bestimmen. Eine der bekanntesten frühen XAI-Methoden dieser Art ist LIME (Ribeiro et al., 2016). LIME ist ein sogenannter modellagnostischer Ansatz, der für verschiedene Arten von Daten – Bilder,

Texte, Tabellen – anwendbar ist. Die Methode basiert darauf, dass für die aktuelle Eingabe eine Menge von sogenannten perturbierten Varianten erzeugt wird, bei denen jeweils Teile der Information gelöscht werden. Die perturbierten Beispiele werden ins Modell eingegeben und darüber identifiziert, welche Informationen vorhanden sein müssen, damit eine bestimmte Klasse ausgegeben wird. Für Bildeingaben werden Pixelgruppen zu sogenannten Superpixeln zusammengefasst, die gemeinsam gelöscht werden.

Die Bezeichnung „Erklärbare KI" ist etwas irreführend: Teilweise wurde der Begriff in der Öffentlichkeit so verstanden, dass es darum geht, die Funktionsweise von KI-Systemen zu erklären. Entsprechend wird inzwischen alternativ die Bezeichnung „erklärend" (*explanatory*) verwendet (Teso & Kersting, 2019; Ai et al., 2021). Erklärbare KI meint auch nicht, dass ein KI-System einen bestimmten Wissensbereich allgemein erklärt, – dies ist die Domäne Intelligenter Tutorsysteme (Polson & Richardson, 2013; Zeller & Schmid, 2016) – sondern dass das Verhalten des Modells allgemein (globale Erklärung) oder für eine bestimmte Eingabe (lokale Erklärung) nachvollziehbar gemacht wird. Viele der ersten Forschungsarbeiten im Bereich XAI haben nicht berücksichtigt, wie komplex die Beziehung zwischen Erklärungen und deren Nutzen für die Nachvollziehbarkeit der Ausgaben von KI-Systemen ist. Es wurde angenommen, dass die Ergänzung einer Modellausgabe durch die Präsentation von relevanten Merkmalen unmittelbar zur Nachvollziehbarkeit führt. Diese naive Herangehensweise wurde insbesondere durch Tim Miller (2019) kritisiert, der aufzeigte, dass XAI notwendigerweise Theorien und empirische Ergebnisse aus der Kognitionswissenschaft und der Mensch-Computer-Interaktion berücksichtigen muss. Zudem stellte sich heraus, dass Erklärungen, die nachvollziehbar machen sollen, warum ein Modell für eine bestimmte Eingabe zu einer bestimmten Ausgabe kommt, nicht immer modelltreu (*faithful*) sind. Das heißt, dass die generierte Erklärung nicht mit dem Prozess übereinstimmt, mit dem das gelernte Modell die Eingabeinformation zur Ausgabe verarbeitet. Beispielsweise konnte für LIME gezeigt werden, dass die als relevant identifizierten Pixel stark variieren, je nachdem, wie diese vom Erklärungsalgorithmus zu Superpixeln gruppiert werden (Schallner et al., 2020). Inzwischen hat sich etabliert, dass Forschungsarbeiten, in denen XAI-Methoden präsentiert werden, eine Evaluation der Modelltreue enthalten müssen.

In den letzten Jahren hat sich die Erkenntnis durchgesetzt, dass Erklärungen nicht per se hilfreich dafür sind, KI-Systeme transparenter zu machen. Zudem gibt es, wie bei Erklärungen durch Menschen, nicht nur eine Art, ein Modell zu erklären. Neben den genannten Relevanzmethoden wurden Methoden für beispielbasierte Erklärungen, kontrafaktische Erklärungen und konzeptbasierte Erklärungen entwickelt (Schwalbe & Finzel, 2024, s. Abb. 10.3). Welche Art von Erklärung hilfreich ist, hängt vom Erklärungskontext ab, also davon, wem was für welchen Informationsbedarf erklärt werden soll (Schmid & Wrede, 2022). Die drei wesentlichen Zielgruppen für Erklärbare KI sind Modellentwickler, Domänenexpertinnen und -experten sowie Endverbrauchende.

Für Modellentwickler ist es besonders relevant, unerwünschtes *overfitting* oder unfaire Verzerrungen im Modell zu identifizieren. Hier sind relevanzbasierte Methoden hilfreich, bei denen für eine Eingabe der Beitrag einzelner Merkmale zur Klassifikation dieser Eingabe aufgezeigt wird. Domänenexpertinnen und -experten benötigen dagegen Informationen, die

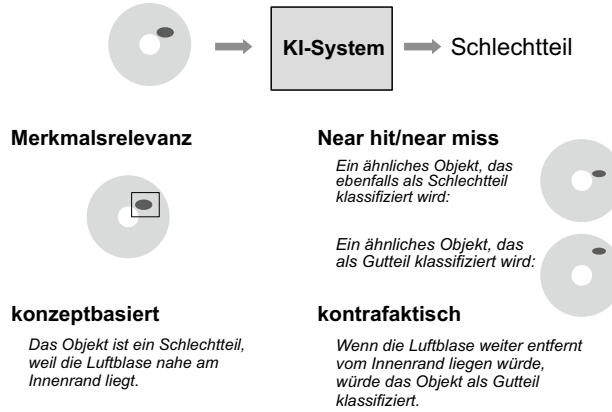

Abb. 10.3 Illustration von verschiedenen Arten von Erklärungen an einem fiktiven Beispiel aus der bildbasierten Qualitätskontrolle

sie dabei unterstützen, die Zuverlässigkeit einer bestimmten Ausgabe besser einzuschätzen, um zu entscheiden, ob sie dem System hier vertrauen oder besser ihrer eigenen, gegebenenfalls abweichenden Beurteilung folgen wollen. Beispielsweise können im Kontext von bildbasierten Diagnosen in der Medizin (Bruckert et al., 2020) konzeptuelle Erklärungen hilfreich sein. Während eine relevanzbasierte Methode lediglich zeigt, dass ein Tumor identifiziert wurde, indem das Tumorgewebe im Eingabebild hervorgehoben wird, ist es für die Einschätzung des Schweregrads des Tumors notwendig, Informationen über die Größe oder die Beziehung zwischen Tumorgewebe und anderen Gewebearten zu berücksichtigen. Sind Konzepte schwer zu benennen, etwa bestimmte Formeigenschaften, bieten sich hier beispielbasierte Erklärungen an (Herchenbach et al., 2022). Erklärungen für Domänenexpertinnen und -experten sollten also dabei helfen, das Vertrauen in ein System sinnvoll zu kalibrieren (s. Abb. 10.1). Beispielbasierte Erklärungen können sich auf die Präsentation von Prototypen (Kim et al., 2016) oder *near hits* und *near misses* (Rabold et al., 2022) beziehen. *Near hits* und *near misse*s sind Beispieleingaben, die der aktuellen Eingabe ähnlich sind und vom Modell zur gleichen Klasse (*near hit*) oder einer anderen Klasse (*near miss*) zugeordnet werden.

Für Endverbrauchende ist es dagegen wichtig zu wissen, ob eine bestimmte Entscheidung, etwa über die Vergabe eines Kredits, die Höhe eines Versicherungsbeitrags oder die Genehmigung einer Fortbildungsmaßnahme, mittels Unterstützung eines KI-Systems erfolgt ist (allgemeine Transparenz). Um nachvollziehbar zu machen, wie eine bestimmte Entscheidung zustande kam, etwa die Ablehnung eines Kredits, können insbesondere kontrafaktische Erklärungen benutzt werden (Wachter et al., 2017). Hier wird bestimmt, welches Merkmal der Eingabe am wenigsten geändert werden müsste, damit das KI-System eine andere Klassenentscheidung treffen würde. Dabei sind nicht änderbare Merkmale ausgeschlossen. Beispielsweise wäre eine hilfreiche Erklärung, dass ein Kredit vergeben würde, wenn die gewünschte Summe zehntausend Euro geringer wäre. Wenig hilfreich wäre eine Erklärung, dass der Kredit vergeben würde, wenn die Person zehn Jahre jünger wäre. Die verschiedenen Arten von Erklärungen sind in Abb. 10.3 mit einem fiktiven Beispiel aus der bildbasierten industriellen Qualitätskontrolle (Herchenbach et al., 2022) illustriert.

Die genannten Arten von Erklärungen können ein KI-System, speziell intransparente tiefe Netze, als *post hoc* Erklärungen ergänzen, um so die Vertrauenswürdigkeitsanforderung der Transparenz zu erfüllen. Alternativ lassen sich direkt interpretierbare Ansätze des maschinellen Lernens nutzen, bei denen die gelernten Modelle in symbolischer Form repräsentiert werden (Atzmueller et al., 2024). Beispiele für solche Modelle sind Entscheidungsbäume, einfache Regressionsmodelle und mit induktiver logischer Programmierung (ILP) gelernte Programme. Interpretierbares maschinelles Lernen wird auch als starkes maschinelles Lernen (Muggleton et al., 2018) bezeichnet. Allerdings sind solche Ansätze nur auf Daten anwendbar, die in symbolischer Form vorliegen. Dies kann in Form von Merkmalen sein, die für Menschen bedeutsam sind (z. B. medizinische Messwerte wie Blutdruck und Cholesterinwert) oder in Form von relationalen Strukturen, etwa chemischen Molekülen. Neurosymbolische Ansätze, bei denen tiefe neuronale Netze zur Klassifikation von komplexen Daten, wie Bilder und regelbasierte Systeme, die etwa mit ILP gelernt werden, kombiniert werden (s. Abb. 10.2b), können auch als Beitrag zur Erklärbaren KI betrachtet werden. Erklärungen tragen also dazu bei, dass die Ausgaben von KI-Systemen nachvollziehbar werden (*explain to understand*). Sie sind aber auch eine Voraussetzung dafür, dass gelernte Modelle durch menschliches Feedback korrigiert und angepasst werden können (*explain to revise*, Finzel et al., 2024).

10.5 Menschliche Kontrolle und Aufsicht

Transparenz ist eine wesentliche Voraussetzung für menschliche Aufsicht von KI-Systemen. Nur wenn nachvollzogen werden kann, aufgrund welcher Information ein KI-System zu einer bestimmten Ausgabe kommt, kann die Zuverlässigkeit und Qualität der Ausgabe beurteilt werden. Menschliche Kontrolle meint zunächst, dass die Entscheidung, sei es eine diagnostische Entscheidung in der Medizin oder die Entscheidung über eine Kreditvergabe, letztendlich immer beim Menschen liegen muss. Allerdings unterliegen Menschen aus verschiedenen Gründen dem sogenannten *automation bias* (Goddard et al., 2012; Gogoll & Uhl, 2018). Teilweise tendieren Menschen dazu, der Ausgabe eines scheinbar objektiven KI-Systems mehr zu vertrauen als ihrer eigenen Einschätzung; teilweise führt Zeitdruck dazu, Ausgaben von KI-Systemen nicht zu hinterfragen. Eine Möglichkeit, solchen Tendenzen entgegenzuwirken, ist die Gestaltung von Mensch-KI-Schnittstellen so, dass die menschliche Entscheidung priorisiert wird. Dies kann zum Beispiel dadurch realisiert werden, dass das KI-System im Hintergrund arbeitet und sich erst einschaltet, wenn die Ausgabe des KI-Systems und die Eingabe einer Entscheidung durch den Menschen voneinander abweichen. Alternativ kann es bereits hilfreich sein, wenn KI-Systeme direkt als partnerschaftliche Systeme (s. Abb. 10.4) konzipiert werden, bei denen Informationen dem KI-System und dem Menschen in geeignet aufbereiteter Form vorliegen. Ebenso sollte den Nutzenden deutlich gemacht werden, dass das KI-System nicht in jedem Fall korrekt, aber dennoch nützlich zur Entscheidungsunterstützung ist. Schließlich können Menschen die Wertigkeit ihrer eigenen Kompetenzen gegenüber dem KI-System

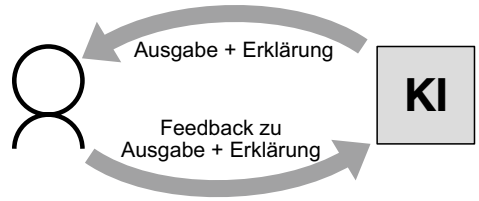

Abb. 10.4 Erklärungen und menschliches Feedback zur Modellanpassung als Zugang zu partnerschaftlichen KI-Systemen

besser wahrnehmen, wenn es die Möglichkeit gibt, dass der Mensch Ausgaben des Systems durch Rückmeldungen korrigieren kann und diese Korrekturen zur Modellrevision genutzt werden (Finzel et al., 2024; Gramelt et al., 2024).

Die Möglichkeit zur Modellrevision durch menschliche Korrektur ist insbesondere auch in Bereichen sinnvoll, in denen zum Zeitpunkt des Modelltrainings die Menge oder die Qualität der Daten nicht genügt. Dies ist einerseits in hochspezialisierten Bereichen der Fall, etwa wenn es um spezielle Krankheitsbilder geht, andererseits immer dann, wenn die Annotation von Trainingsdaten sehr aufwendig ist (Troles et al., 2024). In diesem Fall kann man mit einem noch wenig performanten Modell starten und das Modell inkrementell durch menschliches Feedback verbessern. Des Weiteren bieten sich solche Methoden des interaktiven maschinellen Lernens (Fails & Olsen, 2003) an, wenn KI-Systeme als personalisierte Assistenten eingesetzt werden und sich entsprechend an individuelle Präferenzen anpassen sollen (Göbel et al., 2022). Die Korrektur von Ausgaben, wie Erklärungen eines KI-Systems, ermöglicht neben der expliziten Einbringung von Wissen durch wissensbasierte Systeme (hybride KI, s. Abb. 10.2a) einen zweiten Weg, menschliches Wissen in den Lernprozess einfließen zu lassen: Auch in Bereichen, wo es Menschen schwerfällt, das notwendige Wissen explizit zu formulieren, sind sie oft immerhin in der Lage, fehlerhafte Ausgaben und auch fehlerhafte Erklärungen zu identifizieren und entsprechend zu korrigieren (vgl. Schmid, 2024).

Zentrales Ziel für die Entwicklung partnerschaftlicher KI-Systeme ist es, dass Menschen und KI als gleichberechtigte Partner zusammenarbeiten. Damit können Menschen von der Stärke von KI-Systemen, Muster in sehr komplexen Datenmengen identifizieren zu können, profitieren und gleichzeitig ihre menschlichen Stärken – Erfahrungswissen und Verankerung in einer komplexen, physikalischen Realität – einbringen. Dies gilt speziell für komplexe und sensible Bereiche, in denen es keine Möglichkeit gibt, optimale oder korrekte Entscheidungen zu treffen. Die Gestaltung solcher partnerschaftlicher KI-Systeme ist eine interdisziplinäre Aufgabe, zu der KI-Forschung, Kognitionswissenschaft und Mensch-Computer-Interaktion gemeinsam beitragen müssen.

10.6 Ausblick

Vertrauenswürdige Künstliche Intelligenz umfasst technische und nichttechnische Anforderungen, die ein KI-System erfüllen muss, damit es als vertrauenswürdig gelten kann und Menschen dem KI-System vertrauen können. Performanz und Robustheit, Dis-

kriminierungsfreiheit, Transparenz und Erklärbarkeit sowie menschliche Kontrolle und Aufsicht sind dabei die Anforderungen, die zumindest zu einem großen Teil technisch gelöst werden können. Solche Systeme entsprechen dem von Donald Michie (1988) vorgeschlagenen Konzept des ultrastarken maschinellen Lernens: Das gelernte Modell muss in der Lage sein, das, was es gelernt hat, so an Menschen zu kommunizieren, dass die menschliche Performanz dadurch besser ist, als wenn Menschen nur von den Daten selbst Kenntnis hätten (s. auch Muggleton et al., 2018).

Methoden zur Umsetzung der Anforderungen an Vertrauenswürdige KI wurden bislang vor allem für Klassifikationssysteme entwickelt. Für generative KI besteht hier noch großer Forschungsbedarf. Es gibt keine etablierte Methodik, mit der die Qualität generierter Inhalte automatisch beurteilt werden kann. Hier liegt die Verantwortung allein beim Menschen, die generierten Inhalte kritisch zu reflektieren und gegebenenfalls zu korrigieren. Auch hier wäre das Ziel die Entwicklung von geeigneten Schnittstellen für partnerschaftliche KI-Systeme, so dass das generierte Endprodukt als Co-Kreation zwischen Menschen und KI-System entsteht.

Dieser Beitrag wurde ohne Unterstützung eines generativen KI-Systems erstellt. Auswahl der Inhalte, Strukturierung und Formulierungen stammen ausschließlich von der Autorin selbst.

Danksagung Dieser Beitrag entstand im Zusammenhang mit folgenden drittmittelgeförderten Projekten: dem vom Bayerischen Forschungsinstitut für Digitale Transformation (bidt) geförderten Projekt Mensch-KI-Co-Creation von Programmcode bei unterschiedlichen Vorkenntnissen: Effekte auf Performanz und Vertrauen (pAIrProg, 2024–2028) im Forschungsschwerpunkt Mensch und generative Künstliche Intelligenz: Trust in Co-Creation, und dem vom BMBF geförderten Projekt Ethische Implikationen hybrider Teams aus Mensch und KI-System (Ethyde, Förderkennzeichen 01IS24067B, 2024–2026). Ich bedanke mit herzlich bei Eda Ismail-Tsaous, Sonja Niemann und Celine Spannagl für die kritische Durchsicht des Manuskripts und bei Felix Haase für die Umsetzung der Graphiken.

Literatur

Ai, L., et al. (2021). Beneficial and harmful explanatory machine learning. *Machine Learning, 110*, 695–721.

Atzmueller, M., et al. (2024). Explainable and interpretable machine learning and data mining. *Data Mining and Knowledge Discovery, 38*(5), 2571–2595.

Balzert, H. (1998). *Software-Qualitätssicherung. Lehrbuch der Software-Technik*. Spektrum.

Brinker, T. J., et al. (2019). Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task. *European Journal of Cancer, 113*, 47–54.

Bruckert, S., et al. (2020). The next generation of medical decision support: A roadmap toward transparent expert companions. *Frontiers in Artificial Intelligence, 3*, 507973.

Chew, R., et al. (2019). SMART: An open source data labeling platform for supervised learning. *Journal of Machine Learning Research, 20*(82), 1–5.
Creed, W. D., et al. (1996). Trust in organizations. *Trust in organizations: Frontiers of theory and research, S., 16*, 38.
Dietterich, T. (1995). Overfitting and undercomputing in machine learning. *ACM Computing Surveys (CSUR), 27*(3), 326–327.
Fails, J. A., & Olsen, D. R., Jr. (2003). Interactive machine learning. In *Proceedings of the 8th international conference on intelligent user interfaces (IUI)* (S. 39–45).
Finzel, B., et al. (2024). Near hit and near miss example explanations for model revision in binary image classification. In *International conference on intelligent data engineering and automated learning* (S. 260–271). Springer Nature Schweiz.
Freiesleben, T., & Grote, T. (2023). Beyond generalization: A theory of robustness in machine learning. *Synthese, 202*(4), 109.
Gao, J., et al. (2024). A taxonomy for human-LLM interaction modes: An initial exploration. In *Extended abstracts of the CHI conference on human factors in computing systems* (S. 1–11).
Gao, Y., et al. (2023). Retrieval-augmented generation for large language models: A survey. *arXiv preprint, arXiv*, 2312.10997.
Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals, 14*(2), 627–660.
Göbel, K., et al. (2022). Explanatory machine learning for justified trust in human-AI collaboration: Experiments on file deletion recommendations. *Frontiers in Artificial Intelligence, 5*, 919534.
Goddard, K., et al. (2012). Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association, 19*(1), 121–127.
Gogoll, J., & Uhl, M. (2018). Rage against the machine: Automation in the moral domain. *Journal of Behavioral and Experimental Economics, S., 74*, 97–103.
Gramelt, et al. (2024). Interactive explainable anomaly detection for industrial settings. *arXiv preprint, arXiv*, 2410.12817.
Gunning, D., & Aha, D. (2019). DARPA's explainable artificial intelligence (XAI) program. *AI Magazine, 40*(2), 44–58.
Hall, M. A., et al. (2001). Trust in physicians and medical institutions: what is it, can it be measured, and does it matter? *The Milbank Quarterly, 79*(4), 613–639.
HEG-KI – Hochrangige Expertengruppe für KI der Europäischen Kommission. (2019). *Ethik-Leitlinien für eine Vertrauenswürdige KI*. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai. Zugegriffen am 10.01.2025.
Heidrich, L., et al. (2023). FairCaipi: A combination of explanatory interactive and fair machine learning for human and machine bias reduction. *Machine Learning and Knowledge Extraction, 5*(4), 1519–1538.
Herbold, S., et al. (2023). A large-scale comparison of human-written versus ChatGPT-generated essays. *Scientific Reports, 13*(1), 18617.
Herchenbach, M., et al. (2022). Explaining image classifications with near misses, near hits and prototypes: Supporting domain experts in understanding decision boundaries. In *International Conference on Pattern Recognition and Artificial Intelligence (ICPRAI)* (S. 419–430). Springer International Publishing.
Holliday, D., et al. (2016). User trust in intelligent systems: A journey over time. In *Proceedings of the 21st international conference on intelligent user interfaces* (S. 164–168).
Ilievski, F., et al. (2024). Aligning generalisation between humans and machines. *arXiv preprint, arXiv*, 2411.15626.
Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems, 33*(1), 1–33.

Kaplan, A. D., et al. (2023). Trust in artificial intelligence: Meta-analytic findings. *Human Factors, 65*(2), 337–359.

Kaur, D., et al. (2022). Trustworthy artificial intelligence: a review. *ACM Computing Surveys (CSUR), 55*(2), 1–38.

Kim, B., et al. (2016). Examples are not enough, learn to criticize! criticism for interpretability. *Advances in Neural Information Processing Systems, 29*(NeurIPS 2016), 2280–2288.

Kleinman, Z. (2024). *Why Google's 'woke' AI problem won't be an easy fix*. https://www.bbc.com/news/technology-68412620. Zugegriffen am 10.01.2024.

Krizhevsky, A., et al. (2012a). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems, 25*(NeurIPS 2012), 1097–1105.

Krizhevsky, A., et al. (2012b). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems, 26*(NeurIPS 2013), 1106–1114.

Lapuschkin, S., et al. (2019). Unmasking Clever Hans predictors and assessing what machines really learn. *Nature Communications, 10*(1), 1096.

Lee, N. T. (2018). Detecting racial bias in algorithms and machine learning. *Journal of Information, Communication and Ethics in Society, 16*(3), 252–260.

Lindholm, A., et al. (2022). *Machine learning: A first course for engineers and scientists*. Cambridge University Press.

Maloof, M. A. (2003). Learning when data sets are imbalanced and when costs are unequal and unknown. *ICML-2003 workshop on learning from imbalanced data sets II, 2*, 1–8.

Manhaeve, R., et al. (2021). Neural probabilistic logic programming in DeepProbLog. *Artificial Intelligence, 298*, 103504.

Marra, G., et al. (2024). From statistical relational to neurosymbolic artificial intelligence: A survey. *Artificial Intelligence*, 104062.

Mcknight, D. H., et al. (2011). Trust in a specific technology: An investigation of its components and measures. *ACM Transactions on Management Information Systems (TMIS), 2*(2), 1–25.

Michie, D. (1988). Machine learning in the next five years. In *Proceedings of the Third European working session on learning* (S. 107–122). Pitman.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence, 267*, 1–38.

Mizrahi, M., et al. (2024). State of what art? A call for multi-prompt LLM evaluation. *Transactions of the Association for Computational Linguistics, 12*, 933–949.

Muggleton, S. H., et al. (2018). Ultra-strong machine learning: Comprehensibility of programs learned with ILP. *Machine Learning, 107*, 1119–1140.

Pahl, J., et al. (2022). Female, white, 27? Bias evaluation on data and algorithms for affect recognition in faces. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (S. 973–987).

Pan, S., et al. (2024). Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*.

Polson, M. C., & Richardson, J. J. (2013). *Foundations of intelligent tutoring systems*. Psychology Press.

Rabold, J., et al. (2020). Enriching visual with verbal explanations for relational concepts – Combining LIME with Aleph. In *Machine learning and knowledge discovery in databases: International workshops of ECML PKDD 2019, Würzburg, Deutschland, 16.–20. September 2019, Proceedings, Teil I* (S. 180–192). Springer International Publishing.

Rabold, J., et al. (2022). Generating contrastive explanations for inductive logic programming based on a near miss approach. *Machine Learning, 111*(5), 1799–1820.

Ribeiro, M. T., et al. (2016). 'Why should I trust you?' Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (S. 1135–1144).

Rice, L., et al. (2020). Overfitting in adversarially robust deep learning. In *International conference on machine learning* (S. 8093–8104). PMLR.

Rich, E. (1983). *Artificial Intelligence*. McGraw-Hill.

Robbins, B. G. (2016). What is trust? A multidisciplinary review, critique, and synthesis. *Sociology Compass, 10*(10), 972–986.

Ruggieri, S., et al. (2023). Can we trust fair-AI? In *Proceedings of the AAAI conference on artificial intelligence* (Bd. 37, Nr. 13, S. 15421–15430).

Sarker, M. K., et al. (2022). Neuro-symbolic artificial intelligence: Current trends. *AI Communications, 34*(3), 197–209.

Schallner, L., et al. (2020). Effect of superpixel aggregation on explanations in lime – A case study with biological data. In *Machine learning and knowledge discovery in databases: International workshops of ECML PKDD 2019, Proceedings, Teil I* (S. 147–158). Springer International Publishing.

Schepman, A., & Rodway, P. (2023). The General Attitudes towards Artificial Intelligence Scale (GAAIS): Confirmatory validation and associations with personality, corporate distrust, and general trust. *International Journal of Human-Computer Interaction, 39*(13), 2724–2741.

Schmid, U. (2021). Interactive learning with mutual explanations in relational domains. In S. Muggleton & N. Chater (Hrsg.), *Human-like machine intelligence* (Bd. Kap. 17, S. 338–354). Oxford University Press.

Schmid, U. (2022). Vertrauenswürdige Künstliche Intelligenz. In F. Rostalski (Hrsg.), *Künstliche Intelligenz: Wie gelingt eine vertrauenswürdige Verwendung in Deutschland und Europa?* (S. 287–298). Mohr Siebeck.

Schmid, U. (2024). Trustworthy artificial intelligence – Comprehensible, transparent, correctable. In H. Werthner et al. (Hrsg.), *Introduction to digital humanism* (S. 151–164). Springer.

Schmid, U., & Wrede, B. (2022). What is missing in XAI so far? An interdisciplinary perspective. *KI – Künstliche Intelligenz, 36*(3), 303–315.

Schramm, S., et al. (2023). Comprehensible artificial intelligence on knowledge graphs: A survey. *Journal of Web Semantics, 79*, 100806.

Schwalbe, G., & Finzel, B. (2024). A comprehensive taxonomy for explainable artificial intelligence: A systematic survey of surveys on methods and concepts. *Data Mining and Knowledge Discovery, 38*(5), 3043–3101.

Shanahan, M. (2024). Talking about large language models. *Communications of the ACM, 67*(2), 68–79.

Sheridan, T. B. (2019). Individual differences in attributes of trust in automation: Measurement and application to system design. *Frontiers in Psychology, 10*, 1117.

Stahl, B. C., et al. (2022). Unfair and illegal discrimination. In *Ethics of artificial intelligence: Case studies and options for addressing ethical challenges* (S. 9–23). Springer International Publishing.

Teso, S., & Kersting, K. (2019). Explanatory interactive machine learning. In *Proceedings of the 2019 AAAI/ACM conference on AI, ethics, and society* (S. 239–245).

Thaler, A., & Schmid, U. (2021). Explaining machine learned relational concepts in visual domains-effects of perceived accuracy on joint performance and trust. In *Proceedings of the annual meeting of the cognitive science society* (Bd. 43, Nr. 43, S. 1705–1711).

Tomsett, R., et al. (2020). Rapid trust calibration through interpretable and uncertainty-aware AI. *Patterns, 1*(4).

Troles, J. D., & Schmid, U. (2021). Extending challenge sets to uncover gender bias in machine translation: Impact of stereotypical verbs and adjectives. *Proceedings of the sixth conference on machine translation, WMT@EMNLP, 202*, 531–541.

Troles, J. D., et al. (2024). BAMFORESTS: Bamberg benchmark forest dataset of individual tree crowns in very-high-resolution UAV images. *Remote Sensing, 16*(11), 1935.

Vallor, S. (2024). *The AI Mirror: How to reclaim our humanity in an age of machine thinking*. Oxford University Press.

Wachter, S., et al. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology, 31*(2), 841.

Williams, A., et al. (2022). The exploited labor behind artificial intelligence. *Noema Magazine*, 22. https://www.noemamag.com/the-exploited-labor-behind-artificial-intelligence/. Zugegriffen am 10.01.2025.

Zeller, C., & Schmid, U. (2016). Automatic generation of analogous problems to help resolving misconceptions in an intelligent tutor system for written subtraction. In *Proceedings of the ICCBR Workshops* (S. 108–117).

Zhang, Y., et al. (2020). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (S. 295–305).

Zowghi, D., & Gervasi, V. (2002). The three Cs of requirements: Consistency, completeness, and correctness. In *International workshop on requirements engineering: Foundations for software quality. Essener Informatik Beiträge* (S. 155–164).