**Ute Schmid**

# Künstliche Intelligenz und menschliche Kontrolle

Mit der Entwicklung von hochperformanten Methoden und Architekturen im Bereich maschinelles Lernen eröffnen sich Potentiale für den Einsatz von KI-Methoden für immer mehr Anwendungsbereiche. Inzwischen sind KI-Chatbots in Web-Browser oder Office-Software integriert. Durch diese niedrigschwellige Verfügbarkeit nutzen Menschen in Alltag und Beruf KI-Zusammenfassungen oder lassen sich Präsentationen generieren. Aber auch in komplexen, hoch spezialisierten Bereichen bieten KI-Methoden viele Chancen für mehr Effizienz und Innovation, beispielsweise für genauere Diagnostik in der Medizin oder für die Entwicklung neuer Wirkstoffe. Während in einigen Bereichen voll autonome KI-Systeme entwickelt werden, etwa autonomes Fahren, werden in der überwiegenden Zahl von Anwendungsfeldern Systeme zum Einsatz kommen, bei denen die letztendliche Entscheidung beim Menschen liegen soll oder muss. Dies gilt für Klassifikationssysteme ebenso wie für generative KI-Systeme. Beispielsweise kann Hautkrebs-Screening durch Bildklassifikatoren erfolgen, die diagnostische Entscheidung sowie die Auswahl einer geeigneten Therapie sollte allerdings beim medizinischen Fachpersonal verbleiben. Ein Arztbrief, der mit einem großen Sprachmodell generiert wurde, sollte von Menschen überprüft und gegebenenfalls korrigiert werden. Die Erzeugung von Programmcode für komplexe Anwendungen oder die Nutzung von generativer KI für Scientific Discovery Aufgaben basiert auf ko-kreativen Prozessen, bei denen hohe menschliche Expertise für Instruktion wie Bewertung unverzichtbar ist.

Je mehr kognitive Aufgaben wir an KI-Systeme delegieren, desto größer ist aber die Gefahr, dass wir wesentliche Kompetenzen verlernen oder gar nicht erst erwerben. Um diesem Effizienz-Kompetenz-Paradox sinnvoll entgegenzuwirken, benötigt es neue Methoden zur Gestaltung von Mensch-KI-Schnittstellen, bei denen Menschen nicht zu Abnickern degradiert werden, sondern bei denen menschliche Expertise aktiv nachgefragt ist. Menschliche Kontrolle und Aufsicht werden auch im Kontext der EU-KI-Verordnung, im Kontext der Ethik-Leitlinien für vertrauenswürdige KI als wesentliche Anforderung genannt.

Prof. Dr. Ute Schmid
Lehrstuhl für Kognitive Systeme
Otto-Friedrich-Universität Bamberg
An der Weberei 5, 96047 Bamberg
ute.schmid@uni-bamberg.de
https://www.uni-bamberg.de/en/cogsys/team/schmid-ute/

DOI-Nr.: 10.26125/qh7x-t098

## Warum menschliche Kontrolle bei KI-Systemen unverzichtbar ist

KI-Methoden ermöglichen die maschinelle Lösung von Problemen, die mit Standard-Algorithmen nicht lösbar sind. Insbesondere ermöglichen Methoden des maschinellen Lernens Aufgaben mit dem Computer zu lösen, für die es gar nicht möglich ist, einen Algorithmus explizit zu formulieren. Dies ist in Bereichen der Fall, bei denen Menschen überwiegend implizites Wissen haben. So ist perzeptuelles Wissen nur teilweise oder nicht explizierbar: Es ist kaum möglich, vollständig zu beschreiben, wie man auf beliebigen Bildern erkennt, ob eine Katze abgebildet ist oder aufgrund welcher Eigenschaften man eine Hautveränderung als Melanom identifiziert. Methoden des maschinellen Lernens ermöglichen es, komplexe Muster in großen Mengen von Daten zu identifizieren und darüber zu generalisieren. Die so gelernten Modelle nehmen den Platz von expliziten Programmen ein, indem sie ermöglichen, für neue Eingaben entsprechende Ausgaben zu berechnen. Die gelernten Modelle sind, insbesondere wenn sie auf aktuellen Deep Learning Architekturen basieren, jedoch intransparent (black boxes). Das heißt, es ist nicht nachvollziehbar, aufgrund welcher Information in der Eingabe ein Modell zu seiner Ausgabe kommt.

Im Gegensatz zu KI-Modellen sind bei Standard-Algorithmen die Schritte, in denen Eingaben zu Ausgaben verarbeitet werden, eindeutig festgelegt. Programme – also die konkrete Umsetzung von Algorithmen in einer Programmiersprache – können von Menschen inspiziert, systematisch getestet und für sicherheitskritische Aspekte auch formal verifiziert werden. Auch wenn das bei komplexer Software nicht immer gelingt, kann im Prinzip gewährleistet werden, dass die Programme korrekt und vollständig sind, also dass für jede mögliche Eingabe die passende Ausgabe erzeugt wird. Beispielsweise wird vom Kontrollprogramm eines AirBags erwartet, dass dieser sich bei starkem Aufprall immer öffnet, aber ansonsten nicht. Im Gegensatz dazu kann bei KI-Algorithmen grundsätzlich keine Garantie geben, dass ein Problem korrekt gelöst wird. Wenn wir in einer Suchmaschine nach Bildern suchen, zum Beispiel von Katze auf rotem Sofa, dann erhalten wir meist zwar viele Treffer, aber auch ein paar Bilder, auf denen etwas anderes zu sehen ist. Es ist trotzdem sehr nützlich, dass wir mit ein paar Stichworten aus der riesigen Menge an Daten im Internet gewünschte Inhalte herausfischen können. Wenn der zugrundeliegende KI-Algorithmus bei einem von 100 Fällen einen Fehler macht, ist das nicht weiter schlimm, wir sehen das ja.

Aber was ist, wenn es nicht um Katzenbilder, sondern um Krebsdiagnose, Arbeitslosengeld oder Hochwasserprognose

geht? Hier ist menschliche Kontrolle und Aufsicht entscheidend. Im Gegensatz zu taxonomischem Wissen in der Biologie gibt es bei den genannten Beispielen keine korrekte, sondern nur eine mehr oder weniger plausible Entscheidung. Für die Klassifikation von biologischen Arten gibt es eine sogenannte *ground truth*. In der medizinische Diagnostik wird dagegen ein Goldstandard zugrunde gelegt, also die derzeit zuverlässigste wissenschaftlich abgesicherte Methode. Bei der Entscheidung, ob eine Person Arbeitslosengeld erhalten soll, müssen rechtliche Vorschriften geprüft werden, dabei fließen aber auch subjektive Einschätzungen der Person ein, die den Antrag bearbeitet. Hochwasserprognose basiert auf der Auswertung von Echtzeitdaten, aber auch auf Erfahrungswissen aus vorangegangenen Verläufen für spezielle Gebiete. Bei der Nutzung von KI-Methoden in den genannten Bereichen braucht es beim menschlichen Gegenüber entsprechend fachliche Kompetenz, um die Plausibilität der Ausgabe zu beurteilen. Fehlende Kompetenz, Zeitdruck, aber auch die Neigung von Menschen, Empfehlungen von KI-Systemen blind zu vertrauen (*automation bias*) können hier dazu führen, dass Ausgaben nicht kritisch geprüft, sondern einfach übernommen werden.

Da Algorithmen des maschinellen Lernens über Daten generalisieren, also induktive Schlüsse vom Speziellen, den Trainingsdaten, aufs Allgemeine, ein Modell, das auf neue, unbekannte Eingaben angewendet wird, durchführen, besteht generell keine Korrektheitsgarantie. In bestimmten Bereichen erreichen solche Modelle sehr hohe Genauigkeiten, aber dennoch können Fehler passieren. Aufsicht und Kontrolle kann damit sehr fordernd für Menschen werden. Wenn man vom „Macher“ zum „Überwacher“ wird, ist es nicht einfach, dennoch immer aufmerksam zu bleiben, insbesondere, wenn Fehler sehr selten sind, aber nicht vorherzusehen ist, wann sie passieren.

## Vertrauenswürdigkeit von KI-Systemen

Für die Zusammenarbeit von Mensch und KI-System in komplexen Entscheidungs- und Problemlösungszenarien spielt das Vertrauen, das Menschen KI-Systemen entgegenbringen, eine entscheidende Rolle. Im Kontext der Vorbereitung der Verordnung über Künstliche Intelligenz hat die Expertengruppe für KI der Europäischen Kommission Ethik-Leitlinien formuliert, bei denen der Begriff der Vertrauenswürdigkeit im Zentrum steht. In den Leitlinien wurde konstatiert, dass den vielfältigen Chancen, die durch die Nutzung von KI-Systemen entstehen, 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. Konkret wurden sieben Anforderungen identifiziert: Vorrang menschlichen Handelns und menschliche Aufsicht, technische Robustheit und Sicherheit, Schutz der Privatsphäre und das Datenqualitätsmanagement, Transparenz, Vielfalt, Nichtdiskriminierung und Fairness, gesellschaftliches und ökologisches Wohlergehen und Rechenschaftspflicht.

Die Anforderung an menschliche Kontrolle und Aufsicht wird insbesondere durch Methoden der Erklärbaren KI, kurz XAI für eXPlainable AI, adressiert. Die Entwicklung von XAI-Methoden ist ein noch relativ junges Forschungsfeld, das sich in Reaktion auf die immer komplexeren KI-Architekturen und deren Einsatz in immer mehr Anwendungsbereichen etabliert hat. XAI-Methoden ermöglichen es, Nutzenden von KI-Systemen möglichst gut nachvollziehbar zu machen, aufgrund welcher Information ein (Black Box) Modell zu einer bestimmten Ausgabe kommt. Solche Methoden werden vor allem für sogenannte diskriminative KI-Systeme entwickelt, beispielsweise für Bildklassifikatoren. Im Kontext von bildbasierter medizinischer Diagnostik kann die Ausgabe eines Diagnosevorschlags entsprechend durch eine Erklärung ergänzt werden. Das kann das Hervorheben relevanter Bildteile sein, eine sprachliche Erklärung der als relevant identifizierten Konzepte, etwa Größe, Form oder Lage einer Läsion oder ähnliche Aufnahmen, die vom KI-System gleichartig oder abweichend klassifiziert werden. Diese verschiedenen Arten von Erklärungen ermöglichen es, dass die jeweils speziellen Informationsbedarfe von Nutzenden berücksichtigt werden können. Methoden der Erklärbaren KI sind grundlegend dafür, dass menschliche Kontrolle und Aufsicht von KI-Systemen ermöglicht werden können. Im Kontext von großen Sprachmodellen, wie sie hinter KI-Chatbots stecken und für agentische KI-Systeme sind solche Erklärungsmethoden allerdings weniger relevant. Hier haben Nutzende üblicherweise andere Fragen, beispielsweise, welche Datenquellen in eine generierte Antwort eingeflossen sind oder wie sehr man

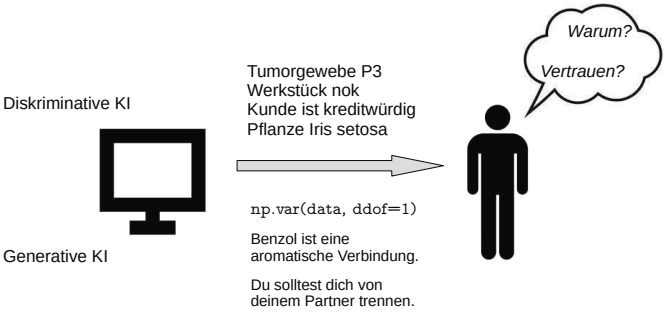



**Abb. 1:** KI-Systeme können in vielen Bereichen in Beruf und Alltag unterstützen. Menschliche Kontrolle und Aufsicht ist dabei aber unverzichtbar. Für ein sinnvoll kalibriertes Vertrauen in die Ausgaben von KI-Systemen sollte nachvollziehbar sein, aufgrund welcher Informationen ein KI-System zu seiner Ausgabe kommt.

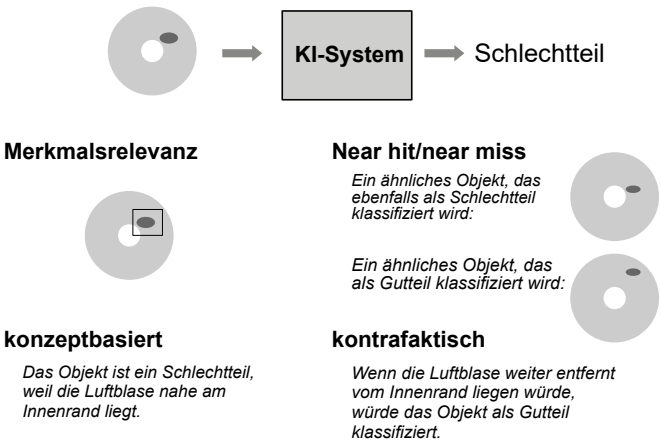



**Abb. 2:** Methoden der Erklärbaren KI (XAI) ermöglichen die Nachvollziehbarkeit der Ausgaben von KI Systemen. Verschiedene Arten von Erklärungen adressieren verschiedene Informationsbedarfe von Nutzenden. (entnommen aus: Schmid, 2026)

darauf vertrauen kann, dass eine Antwort korrekt ist. Hier macht es übrigens keinen Sinn, das Sprachmodell direkt zu fragen, wie sicher es sich ist. Unsicherheitsabschätzungen basieren auf komplexen mathematischen Methoden, bezogen auf die interne Verarbeitung von Informationen im Sprachmodell.

## Konsequenzen für die Bildung

Die empirischen Befunde mehren sich, bei denen gezeigt wird, dass eine übermäßige Delegation von Aufgaben an KI-Chatbots dazu führt, dass wesentliche Kompetenzen nicht erworben werden. Um die Ausgaben von KI-Systemen kritisch zu bewerten, ist es nötig, selbst über das entsprechende fachliche Wissen zu verfügen. Nur wenn ich etwas selber sehr gut kann, kann ich ein KI-System sinnvoll prompten und die Ausgabe auf Korrektheit oder Qualität bewerten. Fachliche Kompetenz und Bewertungskompetenz erwirbt man nur dadurch, wenn man Aufgaben selber löst (learning by doing). Die Befunde der neuen PISA-Studie zeigen Defizite im Leseverständnis, bei Mathematik und in den Naturwissenschaften. Besonders kritisch sehe ich, dass in Deutschland der schulische Erfolg sehr stark vom Elternhaus abhängt. Hier ist sicher ein Faktor, dass das so wichtige individuelle Üben sehr stark auf die Hausaufgaben und damit ins Elternhaus verlegt ist. Meiner Meinung nach brauchen wir ein gutes Ganztagskonzept, bei dem individuelles Üben im schulischen Kontext stattfindet. Dies kann durchaus sinnvoll durch KI-Systeme, speziell Intelligente Tutorsysteme, unterstützt werden, bei denen beim Lösen von Aufgaben gezielt Missverständnisse identifiziert werden, die fehlerhaften Lösungen zugrunde liegen und geeignetes didaktisches Feedback gegeben wird. Statt Lehrpläne zu überladen, sollte mehr Platz für zentrale fachliche Kompetenzen sein, die dann vertieft vermittelt werden können, um auch Verständnis von Zusammenhängen, Problemlösetransfer und Reflexion zu fördern.

## Referenzen

[1] Atzmueller, M., Fürnkranz, J., Kliegr, T., & Schmid, U. (2024). Explainable and interpretable machine learning and data mining: M. Atzmueller et al. *Data Mining and Knowledge Discovery*, **38(5)**, 2571-2595. https://doi.org/10.1007/s10618-024-01041-y

[2] Braun, T., Hüllermeier, E., Schmid, U., & Görz, G. (2026). Handbuch der Künstlichen Intelligenz (7. Auflage). De Gruyter. https://www.degruyterbrill.com/de/document/doi/10.1515/9783111386690/html

[3] 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 (Abruf 10.01.2025)

[4] Schmid, U. (2026). Vertrauenswürdige Künstliche Intelligenz. In Schmiedchen, F. u.a. (Hrsg.): Künstliche Intelligenz und Wir – Stand, Nutzung und Herausforderungen der KI. Springer. https://link.springer.com/chapter/10.1007/978-3-662-71567-3_10

[5] Schmid, U. (2025). KI Kompetenzen (AI Literacy). In Brägger, G. & Rolff H.-G. (Hrsg.): Handbuch Lernen mit digitalen Medien. Beltz. https://beltz.de/fachmedien/paedagogik/handbuch-lernen-mit-digitalen-medien/BEL83244

**Prof. Dr. Ute Schmid**



© Jürgen Schabel

Ute Schmid ist Inhaberin des Lehrstuhls für Kognitive Systeme an der Universität Bamberg. Sie ist geschäftsführende Direktorin des Bamberger Zentrums für Künstliche Intelligenz (BaCAI). Sie hat an der TU Berlin studiert und Diplom-Abschlüsse in Informatik und Psychologie. Ebenfalls an der TU Berlin hat sie in Informatik promoviert (1994) und habilitiert (2001). Forschungsschwerpunkte sind erklärbare KI, neurosymbolisches maschinelles Lernen mit Anwendungen in der bildbasierten Diagnostik und in der Bildung. Ute Schmid Mitglied im Direktorium des Bayerischen Instituts für Digitale Transformation (bidt). Sie ist Mitglied im Bayerischen KI Rat und Sprecherin der AG 1 (Technologien und Data Science) der Plattform für Lernende Systeme. Sie leitet die AG KI in Schulen (KiS) am Fachbereich KI der Gesellschaft für Informatik. Ute Schmid ist Mitglied in zahlreichen wissenschaftlichen Beiräten, darunter im KI Campus sowie bei AI Grid. Für ihre Forschung im Bereich kognitive Künstliche Intelligenz wurde sie 2022 zur EurAI Fellow ernannt. Für ihr Engagement im Bereich KI, Frauen in der Informatik und Informatikbildung wurde sie 2023 zur GI Fellow ernannt. Im Jahr 2025 wurde sie zum Mitglied der Deutschen Akademie für Technikwissenschaften (acatech) gewählt. Sie ist Mitglied der Luxemburgischen Akademie der Wissenschaften und Künste (LASA). Für ihre Aktivitäten in der Wissenschaftskommunikation im Bereich KI wurde sie 2026 mit dem DFG Communicator Preis ausgezeichnet.

