Melanie Platz und Markus Peschel

Kann man Künstlicher Intelligenz vertrauen?

Prof. Dr. Melanie Platz, ist Professorin für Didaktik der Primarstufe – Schwerpunkt Mathematik an der Universität des Saarlandes, Schwerpunkte sind Lernen und Lehren mit und über Medien im Mathematikunterricht, mathematisches Begründen und Beweisen in der Primarstufe.

Prof. Dr. Markus Peschel, ist Professor für Didaktik des Sachunterrichts an der Universität des Saarlandes und Fachreferent für Lernkulturen und Sachunterricht im Grundschulverband. Schwerpunkte sind Digitalisierung, Experimentieren und Lernkulturen.

Insbesondere die Veröffentlichung von ChatGPT 2022 hat Künstliche Intelligenz (KI) in der Gesellschaft in den Fokus gerückt, da KI dadurch für die breite Öffentlichkeit zugänglich gemacht wurde, was eine große Faszination auslöste (Baker 2024); hier sei nur auf die aktuellen Entwicklungen in der EU nach dem Pariser KI-Gipfel (2025) verwiesen. „KI" und „ChatGPT" werden seitdem häufig synonym verwendet. Was aber ist KI?

Ein Ziel der Künstlichen Intelligenz (KI) besteht darin, Intelligenz außerhalb des menschlichen Gehirns zu schaffen (Buxmann/Schmidt 2021), d. h. Maschinen mit Fähigkeiten auszustatten, die dem intelligenten Verhalten von Menschen ähneln (Kaplan/Haenlein 2019). Katharina Zweig (2019) definiert KI folgendermaßen: „Als künstliche Intelligenz bezeichnet man eine Software, mit deren Hilfe ein Computer eine kognitive Tätigkeit ausführt, die normalerweise Menschen erledigen." (S. 126)

KI wird häufig als eine Innovation der Neuzeit wahrgenommen, obwohl es bereits seit der Antike Ideen gibt, künstliches Leben zu erschaffen. Allerdings wurden die heutigen KI-Systeme wie ChatGPT[1] erst durch die Rechenleistung moderner Computer, die Fortschritte bei künstlichen neuronalen Netzen und Deep Learning und das Konzept des ‚Trainierens' von Algorithmen an großen Datenmengen ermöglicht. Und die nötigen ‚Big Data' wurden erst durch das Internet verfügbar (Ertel 2021). Dabei gilt: Je mehr Trainingsdaten zur Verfügung stehen, desto besser wird die Leistung der zugehörigen KI sein (Andree 2023). ‚Big Tech' wie Google, Amazon, Facebook, Apple und Microsoft nutzen riesige Datenmengen aus dem Netz und von ihren Nutzer:innen und können so leistungsfähige KI-Anwendungen entwickeln. Laut AGB kann auch OpenAI jede Information, die in einen ChatGPT-Dialog eingegeben wird, analysieren und für das Training von KI-Modellen nutzen (Baker 2004).

Allerdings garantiert die Größe des Datensatzes keine Diversität (Bender et al. 2021): Diskriminierende Inhalte in den Daten und soziale Ungerechtigkeit können reproduziert und verstärkt werden und zu algorithmischem Bias führen, mit schwerwiegenden Folgen für benachteiligte Gruppen. Von KI-Systemen gehen allerdings noch weitere Gefahren aus wie Diskriminierung, mangelnde Erklärbarkeit und Transparenz, Fehlinformationen, Halluzinationen und Manipulationsgefahr, Abhängigkeit und Verlust menschlicher Kreativität, fehlender Schutz geistigen Eigentums, mangelnde Nachhaltigkeit und einseitige kulturelle Prägung (Bender et al. 2021; Miao/Holmes 2023; Platz/Plote 2025; siehe Abb. 1). Insbesondere die kulturelle Prägung hat deutlichen Einfluss auf die Nutzung und die Lernkultur in der Schule, wie und wozu KI in der Grundschule genutzt wird (Peschel et al. 2023).

Es verwundert nicht, dass das Thema der Altersangemessenheit des Einsatzes von KI diskutiert wird (z. B. Ständige Wissenschaftliche Kommission (SWK) 2024; Kultusministerkonferenz (KMK) 2024). Dennoch sind bereits Kinder von digitalen Medien und KI umgeben und nutzen diese. Die Annahme, dass Kinder heutzutage Digital Natives sind oder zu Hause bereits genügend digitale Kompetenzen erwerben, wurde u. a. von der aktuellen ICILS-Studie (Eickelmann et al. 2024) deutlich revidiert; bereits Irion sprach von den Digital Naives (Irion 2018) und forderte entsprechende curriculare Lernszenarien.

Dabei erinnert eine mögliche Verweigerung, KI in der Grundschule einzusetzen, an die Vermeidungsdiskussionen über Mediennutzung, Computernutzung, Tabletnutzung oder Digitalisierung insgesamt in den 1990er-, 2000er-, 2010er-Jahren, ist aber spätestens seit der KMK-Empfehlung *Bildung in der digitalen Welt* (2016) auch für Grundschulen obsolet und lässt sich auch nicht auf den privaten Bereich auslagern.

Um eine entsprechende „digital literacy"[2] (Peschel 2022) zu erwerben, ist es u. a. wichtig zu wissen, wie man Ausgaben einer KI wie ChatGPT bewertet, und es ist wichtig, Risiken durch die Preisgabe personenbezogener Daten abzuschätzen (Platz 2025; Platz/Plote 2025; GDSU 2021; Gervé et al. 2023).

Nach KMK (2024) sind für das ‚*Lernen über KI*' „neben einer grundlegenden informatischen Bildung insbesondere über KI sowie über ihre Wirkungsweisen – auch die Klärung ihrer ethischen und rechtlichen Rahmenbedingungen bei Lehrkräften, Schülerinnen und Schülern sowie in der Bildungsadministration selbst erforderlich." Künstliche Intelligenz kann insgesamt große Chancen für die Gesellschaft wie auch für den schulischen Bereich bieten. Grundlage ist allerdings ein verantwortungsvoller Umgang mit den Chancen (und Grenzen) der KI. Die Entwicklung solcher Kompetenzen im Sinne einer *digital literacy* erfordert nicht nur das *Lernen mit KI*, sondern auch ein fundiertes *Lernen über KI*, um Kindern einen dauerhaft souveränen und kritischen Umgang mit dieser sich ständig weiterentwickelnden Technologie zu ermöglichen (Platz/Plote 2025; Gesellschaft für Didaktik des Sachunterrichts GDSU 2025).

Wie kann ein solches Lernen über KI aussehen? Es folgt ein Beispiel zu Sprachmodellen.

Lernen über KI am Beispiel von Sprachmodellen

Eine zentrale Frage des Lernens *über* Digitalsierung bzw. *über* KI ist, ob man einer KI vertrauen kann. Die Eloquenz bzw. Sprachgewandtheit von Large Language Models wie ChatGPT führt dazu, dass man sich mit ihnen relativ gut ‚unterhalten' kann, ihre Fähigkeiten nutzt und anfängt, die Technologie zu personifizieren und in gewisser Weise zu vermenschlichen. Um aber einer Technologie vertrauen zu können, sind einerseits Durchschaubarkeit und Erklärbarkeit wichtig, andererseits Genauigkeit und Zuverlässigkeit der Ausgaben/Aussagen. *Durschaubarkeit* bedeutet, dass die Entwickler:innen der Technologie den Anwender:innen so viele Informationen zur Verfügung stellen, dass man als Nutzer:in verstehen kann, wie sie funktioniert und vor allem, wie bzw. auf welcher Basis die Maschine Entscheidungen trifft. Hier sind relativ schnell ethische Abwägungen betroffen, insbesondere, wenn es um Minderheiten- oder Kinderschutz geht. *Erklärbarkeit* bedeutet, dass man als Nutzer:in herausfinden kann, woher eine bestimmte Ausgabe eines Systems kommt. Bei der *Genauigkeit* geht es darum, wie richtig die Ausgaben einer Technologie sind und wie gut die Technologie funktioniert. *Zuverlässigkeit* bedeutet, dass die Technologie keine Ausfälle haben darf.

Es drängt sich die Frage auf, ob man einer KI wie ChatGPT vertrauen kann – das heißt, ob sie durchschaubar, erklärbar, genau und zuverlässig ist. ChatGPT ist ein KI-Chatbot. KI-Chatbots nennt man auch „stochastische Papageien" (Bender et al. 2021), was bedeutet, dass KI-Sprachmodelle zwar gut klingende Sprache erzeugen können, aber die Bedeutung der Sprache und vor allem den Kontext nur begrenzt verstehen (Platz/Schick angenommen).

Alle Sprachmodelle – auch GPT – folgen einer solchen Sprachwahrscheinlichkeitsverteilung (s. Kasten), Genauigkeit und Zuverlässigkeit sind allerdings wesentlich besser als in dem einfachen Beispiel des Bi-Gram-Sprachmodells, da GPT gigantische Textmengen und nicht nur den direkten Vorgänger als Grundlage berücksichtigen und auch kontextualisieren kann. Allerdings ist GPT nahezu undurchschaubar sowie unerklärbar eben aufgrund der undurchschaubaren Trainingsdaten, der Nichterklärbarkeit der Ausgabe sowie von Halluzinationen (Platz/Schick angenommen). Dies führt dazu, dass GPT IMMER Antworten geben wird und diese auch immer präziser bzw. passender werden; allerdings ist nicht sicher, dass die Antworten korrekt sind. Im Rahmen der *Explainable Artificial Intelligence* wird derzeit daran geforscht, wie nachvollziehbar gemacht werden kann, auf welche Weise KI zu Ergebnissen kommt. Die Durchschaubarkeit und die Erklärbarkeit solcher Systeme sollen dadurch erhöht werden.

Oben genanntes Beispiel kann in einem erweiterten fächerübergreifenden Ansatz mit Einbezug des Sachunterrichts aufgegriffen werden (in Anlehnung an Peschel/Platz 2024). Oldenburg (2022) beschreibt die Pole der Digitalisierung in Anlehnung an die Funktionen des Sachrechnens nach Winter (1995): Digitalisierung als Lernstoff, als Lernprinzip und als Lernziel. Übertragen auf KI ergeben sich (Platz/Decker/Plote 2023):
(1) KI als Lerninhalt: Wissen über und Fertigkeiten im Umgang mit KI werden aufgebaut, hier wird KI-Literacy adressiert.
(2) KI als Lernprinzip: Bezüge zur Realität herstellen, um die Schüler:innen für ein Bewusstsein von KI in der Lebenswelt zu sensibilisieren, ihr Verständnis zu fördern und ihre Kenntnisse und Fertigkeiten zu stärken. Bezogen auf KI geht es um die Frage: Wie, wo und warum sind Informationen im Internet repräsentiert und wie muss ich mit der KI kommunizieren, um Erkenntnisse / Wissen für mein Leben zu generieren?
(3) KI als Lernziel: die umfassendste Funktion, in ihr sind die zuvor genannten aufgehoben. Insgesamt handelt es sich um „[…] ein didaktisches Programm, in das tiefere Dimensionen pädagogischen Arbeitens eingehen: die übergeordneten Ziele des Mathematikunterrichts [und Sachunterrichts] (sein möglicher Beitrag zur Entfaltung der Kreativität und zur Sensibilisierung für die Probleme unserer Welt) und das Bild, das man vom Menschen und menschlichem Lernen hat" (Winter 1995, 37). Hier werden Fragen gestellt wie: Sind die Ausgaben der KI vertrauenswürdig und brauchbar? Welche Informationen über mich erhalten Dritte durch meine Eingaben?

Dabei lässt sich (1) vornehmlich im Mathematikunterricht verorten, da viele KI-Grundlagen auf elementarisierbaren mathematischen Konzepten basieren (z. B. Platz im Druck).

(2) lässt sich an der Schnittstelle Mathematik- und Sachunterricht verorten, zum einen, da der Mathematikunterricht zum Mündigwerden durch und gegenüber Mathematik unter Beachtung digitaler Perspektiven beitragen sollte. Schüler:innen und Lehrpersonen können dem Einsatz von KI nur durch mathematische Bildung kompetent begegnen (Bescherer et al. 2024, 10). Zum anderen ist diese Mündigkeit über die algorithmische Beeinflussung von Suchergebnissen und Informationsblasen ein

Ein einfaches Sprachmodell als Beispiel

Um die Wahrscheinlichkeit von Sprachausgabe in ChatGPT nachvollziehen zu können, kann ein einfaches Sprachmodell bereits in der Primarstufe genutzt werden, um in einem transdisziplinären Ansatz der Mathematikdidaktik mit der Deutschdidaktik zu vermitteln, wie ChatGPT im Grunde funktioniert: Mit dem Bi-Gram-Sprachmodell (Platz/Schick angenommen; Platz 2024; Arnold 2023; Jurafsky/Martin 2018; https://www.soekia.ch) kann man beispielsweise die Geschichte „Die Raupe Nimmersatt" als Trainingsdatensatz verwenden. Der Algorithmus funktioniert so, dass die Geschichte abhängig vom vorherigen Wort weitererzählt wird. Das nächste Wort wird mittels Zufallsexperiment abhängig von Worthäufigkeiten bestimmt. Die Ausgabe, also die weitererzählte Geschichte, ergibt in der Regel nicht viel Sinn, da jedes Wort nur vom direkten Vorgänger abhängt. Außerdem können unbekannte Wörter nicht verarbeitet werden (Platz/Schick angenommen). Die meisten Kinder werden zustimmen, dass das Bi-Gram-Sprachmodell zwar durchschaubar und erklärbar ist, allerdings sind die Genauigkeit und die Zuverlässigkeit eingeschränkt.

zentrales Anliegen beim Lernen über Digitalisierung und über KI im Sachunterricht, da die Bewusstheit über die Daten ein zentrales Zukunftsfeld darstellt. Zur Erfüllung dieser Funktion sind allerdings auch die weiteren Lernbereiche und Fächer, beispielsweise Deutsch, von großer Bedeutung.

(3) lässt sich vornehmlich im Sachunterricht als Beitrag zur Lebenswelterschließung verorten, wobei hier insbesondere die Perspektiven- und Fächerverbindung zum Tragen kommt (Peschel/Platz 2024). Die Auswirkungen der Verwendung von KI auf die Lebenswelt der Kinder sind zu thematisieren und es ist zu prüfen, inwiefern aus einem mathematisch-algorithmischen Verständnis heraus das Lernen über den Einfluss von KI auf die eigene Lebenswelt, und den Umgang mit eigenen und fremden Daten sowie ein Vertrauen in die Expertise von Fachinhalten entwickelt werden kann. Dies betrifft im Konkreten auch den Umgang mit Demokratieeinflüssen, Fake News und die Mitgestaltbarkeit der digitalen Welt. Weitere Informationen zu diesem sachunterrichtsdidaktischen Zugang finden sich im Artikel von Peschel et al. in diesem Heft.

Fazit

Wenn wir für unsere Kinder wollen, dass sie kritische KI-Bürger:innen, verantwortungsbewusste Nutzer:innen von KI, Mitgestalter:innen von KI-Werkzeugen und nicht nur passive Nutzer:innen, beeinflusst von Deepfakes, werden (Miao 2024), dann müssen ab der Primarstufe eine *digital literacy* bzw. KI-Kompetenzen aufgebaut werden, um die aktiven Nutzer:innen der nächsten KI-Generationen auszubilden und nicht nur passives Konsumieren zu unterstützen. Dabei geht es nicht nur darum, die Kinder an die digital und KI-geprägte Welt anzupassen, sondern auch darum, sie in die Lage zu versetzen, diese zu analysieren, zu reflektieren und zu gestalten (vgl. auch das RANG-Modell im Glossar).

Die UNESCO hat 2024 entsprechend einen KI-Kompetenzrahmen für Schüler:innen (Miao/Shiohira 2024) und einen für Lehrpersonen (Miao/Cukurova 2024) veröffentlicht. Darin werden vier Aspekte als zentrale KI-Kompetenzen definiert: 1. menschenzentrierte Denkweise, 2. Ethik der KI, 3. KI-Techniken und -Anwendungen und 4. KI-System-Design, deren Erwerb Lernende zu verantwortungsvollen Nutzenden und Mitgestaltenden der digitalen Welt machen und sie auf ihre Rollen in der nächsten KI-Generation vorbereiten sollen. Wichtig im Umgang mit KI ist, stets kritisch die ausgegebenen Informationen zu prüfen, z. B. mit einer alternativen Suchmaschine oder durch Expertenbefragungen/-prüfungen.[3]

Für das ‚Lernen *mit* KI' ist entsprechend ein ‚Lernen *über* KI' wichtig und für den Einsatz im Unterricht ist KI immer auf ethische und schutzwürdige Aspekte zu prüfen[4]. Ziel sollte ein bewusster Umgang mit allen Aspekten der Digitalisierung und – neu – der KI sein, um Kinder zu informierten und mündigen und gestaltenden Subjekten ihrer Lebenswelt auszubilden. Ein Vertrauen zu KI kann nur dann aufgebaut werden, wenn sie durchschaubar, erklärbar, genau und zuverlässig ist. Der *EU AI Act* betont die Risiken, die von KI ausgehen können, und zielt darauf ab, Innovationen mit Sicherheits- und Ethikstandards in Einklang zu bringen und die Entwicklung vertrauenswürdiger KI im Bildungsbereich zu fördern. Diese entstehen derzeit schulartübergreifend in vielen Bundesländern.

Anmerkungen

1 Auch Chatbots hat es schon wesentlich früher gegeben (z.B. ELIZA 1966), allerdings sind diese nicht auf so großes Interesse gestoßen wie ChatGPT (Platz/Plote 2025).
2 Digital Literacy ist die Kompetenz sich mündig in einer Welt, die zunehmend von Digitalisierung und Digitalität geprägt ist, zu bewegen.
3 Da populäre Suchmaschinen ihre Algorithmen nicht offenlegen und auch hier KI zu fragwürdigen/problematischen Ergebnissen führen kann, eignen sich z.B. Suchmaschinen für Kinder wie FragFINN oder Helles Köpchen oder alternative Suchmaschinen für Erwachsene wie DuckDuckGo, Ecosia, Startpage oder Qwant.
4 Beispielsweise die „Ethischen Leitlinien für Lehrkräfte über die Nutzung von KI und Daten für Lehr- und Lernzwecke" der Europäischen Kommission definieren Anforderungen, die KI-Systeme im Bildungsbereich erfüllen müssen (Europäische Kommission 2022).

Literaturangaben zum Artikel können Sie von unserer Website herunterladen: https://t1p.de/GSa170Lit

Ungerechtigkeit Für Lernende mit niedrigem sozio-ökonomischen Hintergrund stellen finanzielle Hürden durch private Anbieter und mangelnde Unterstützungsmöglichkeiten durch die Eltern eine Zugangsbarriere dar (SWK 2024).	**‚Verunreinigte' Trainingsdaten, Bias (Verzerrung) und Diskriminierung** Die Intransparenz von Trainingsdaten wird zunehmend kritisch, da KI-generierte Inhalte das Internet verunreinigen' und realistischere Deepfakes entstehen.	**Mangelnde Erklärbarkeit und Transparenz** Die Funktionsweise von KI-Modellen ist oft intransparent, ihre Algorithmen schwer erklärbar.	**Fehlinformationen, Halluzinationen und Manipulationsgefahr** Dies kann gezielt für Desinformation oder Manipulation missbraucht werden und stellt eine potenzielle Bedrohung für persönliche Sicherheit und demokratische Prozesse dar.
Abhängigkeit, Verlust menschlicher Kreativität Es besteht die Gefahr, dass Lernende und Lehrpersonen zunehmend passiv auf KI-Lösungen zurückgreifen, anstatt aktiv Probleme zu lösen. Dies könnte langfristig die Entwicklung von Kreativität und kritischem Denken beeinträchtigen.	**Schutz geistigen Eigentums** Mechanismen zum Schutz von Urheberrechten oder zur Entschädigung fehlen weitgehend.	**Mangelnde Nachhaltigkeit** KI-Modelle sind äußerst ressourcenintensiv. Sie benötigen viel Energie und die Kühlung der Rechenzentren erfordert enorme Wassermengen, oft in Regionen mit knappen Wasserressourcen.	**Arbeitsbedingungen** Betrieb und Wartung von KI-Modellen basieren oft auf menschlicher Arbeit. Diese Aufgaben werden häufig in Billiglohnländern unter schlechten Bedingungen ausgeführt.
	Kulturelle Perspektiven Es besteht die Gefahr einer einseitigen kulturellen Prägung, da die Inhalte von den Ländern oder Konzernen dominiert werden, die die KI entwickeln und trainieren.	**Geschwindigkeit der Entwicklung von KI** Die Geschwindigkeit der Entwicklung von KI ist schneller als die Anpassung der nationalen Rechtsvorschriften.	

Abb. 1: Kritikpunkte an großen KI-Modellen (Platz/Plote 2025)