Von Marvin Kopka

Gesundheitsrat per Algorithmus: Wie KI berät

„Ich habe schon ChatGPT gefragt, aber was meinst du zu…?“ – dieser Satz taucht inzwischen in vielen Beratungskontexten auf.
Aber darf man alles glauben, was die KI an Antworten liefert? Wie funktioniert KI? Und wie können Eltern erkennen, welche Ratschläge hilfreich oder – im Gegenteil – gefährlich sind?

Warum (werdende) Eltern gerne KI nutzen, ist nachvollziehbar: KI-Chatbots sind rund um die Uhr verfügbar, oft kostenlos und können auf nahezu jede Frage eine Antwort geben. Für das Gesundheitssystem und die Gesundheitskompetenz der Gesamtbevölkerung ist das gleichermaßen eine Chance und eine Herausforderung: Einfache medizinische Fragen lassen sich schnell selbst beantworten (und zwar deutlich personalisierter als durch „Dr. Google“), wodurch bei allen Beteiligten Ressourcen gespart werden können.
Auf der anderen Seite können KI-Ratschläge aber auch die Sicherheit der Nutzer*innen gefährden, wie ein Beispiel aus der Inneren Medizin zeigt: Ein 60-jähriger Mann betrat eine Notaufnahme in Washington und gab an, zu glauben, dass sein Nachbar ihn vergifte (1). Nach ausführlicher Diagnostik konnte eine Bromid-Intoxikation diagnostiziert werden – eine Vergiftung, die im frühen 20. Jahrhundert häufig war, Somnolenzen, Psychosen und Krampfanfälle auslösen kann, durch neue Sicherheitsstandards aber bereits seit längerer Zeit selten auftritt. Hatte der Nachbar den 60-jährigen Mann also wirklich mit Bromsalz vergiftet? Nach Aufnahme in der Psychiatrie und entsprechender Behandlung verbesserte sich der Zustand des Mannes. Er gab an, er habe gesünder leben und deshalb Speisesalz reduzieren wollen. Dafür habe er ChatGPT konsultiert und im Anschluss über drei Monate Speisesalz durch Bromsalz ersetzt – und sich somit selbst vergiftet.
Obwohl dieser Fall vermutlich ein Extrembeispiel ist, verdeutlicht er das Problem mit KI-Tools sehr gut: Während medizinische Laien sich früher selbst mühsam durch verschiedenste Quellen kämpfen und alle Informationen zusammentragen mussten, um eine Entscheidung zu treffen (und bei dieser Recherche hoffentlich auf entsprechende Warnungen gestoßen sind!), wird dieser Prozess heute an ein KI-Tool ausgelagert. Die Aufgabe der Nutzer*innen ist es lediglich noch, zu entscheiden, ob diese Empfehlung angenommen oder abgelehnt wird, das heißt, diese Informationen mit bestehendem Wissen und Erfahrungen abzugleichen (2). Ohne fachliche Expertise kann dieser Abgleich allerdings gefährlich werden, da hier auch die Metakognition relevant ist, also zu wissen, was man nicht weiß.

Das soll nicht heißen, dass KI-Tools grundsätzlich falsch liegen oder gefährlich sind – im Gegenteil, verschiedene Studien zeigen sogar, dass KI-Tools sehr oft richtig liegen und gute Antworten geben (3). Schwierig ist nur, einzuschätzen, wann Empfehlungen richtig oder falsch sind. Was Nutzer*Innen aber einschätzen können, sind die potenziellen Konsequenzen, die aus Fehlern entstehen können. Falsche Wörter oder seltsame Formulierungen in einem Text, den ChatGPT zusammengefasst hat, sind nicht weiter dramatisch. Sich selbst über drei Monate mit Bromsalz zu vergiften hingegen schon.

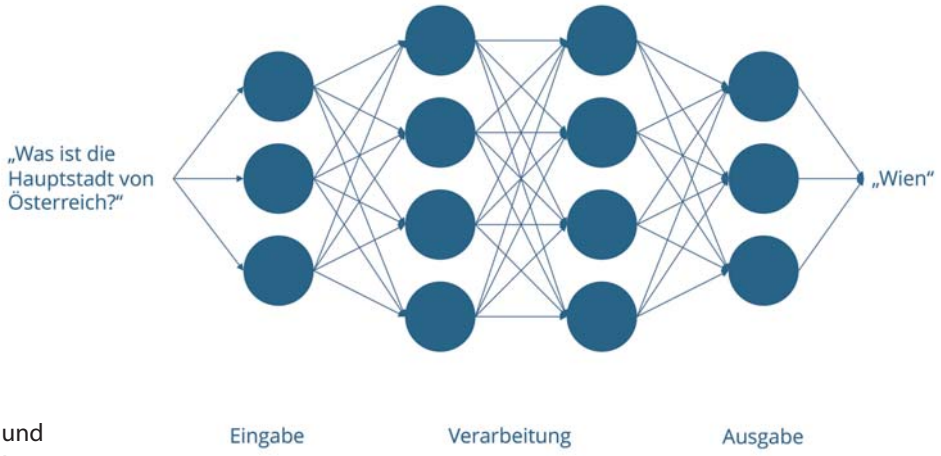


Abb. 1. Vereinfachte Darstellung eines neuronalen Netzes.

Was ist KI und was ist ein Large Language Model?

Um zu verstehen, weshalb ChatGPT und andere KI-Tools überhaupt falsche Antworten generieren, ist es wichtig zu verstehen, was KI überhaupt ist. „Künstliche Intelligenz" ist ein Sammelbegriff, der sehr viele verschiedene mathematische Verfahren zusammenfasst. Am häufigsten wird damit maschinelles Lernen gemeint, also aus Daten Muster abzuleiten und daraus Vorhersagen zu treffen. Das kann mit sehr einfachen Entscheidungsbäumen und Regeln geschehen (zum Beispiel: „Wenn Fieber bei über 38 Grad liegt und das Kind unter drei Monate alt ist, sollte ein Arzt konsultiert werden"), aber auch mit probabilistischen Ansätzen, das heißt, aus vielen verschiedenen Zahlen wird etwas Neues vorhergesagt. Dabei müssen gleiche Zahlen aber nicht zwangsläufig zur selben Vorhersage führen, da es eine Zufallskomponente gibt.

Das zeigt sich besonders in neuronalen Netzen (Abb. 1). Diese sind menschlichen Nervensystemen nachempfunden und bestehen aus sehr, sehr vielen Knoten (Neuronen), die miteinander verbunden sind. Zu Beginn wird eine Eingabe (zum Beispiel eine Frage, die man ChatGPT stellt) in Zahlen übersetzt und diese Zahlen werden an die nächsten Knoten weitergegeben, die wiederum neue Zahlen generieren und auch wieder weitergeben. Das geht immer so weiter, bis eine bestimmte Ausgabe (zum Beispiel eine Antwort auf die gestellte Frage) erstellt wird.

Das klingt einfach, neuronale Netze bestehen aber oft aus mehreren Milliarden solcher Knoten. Was also genau innerhalb dieser neuronalen Netze passiert, ist schwer nachvollziehbar; wir wissen nur, dass hier sehr viele Zahlen miteinander verrechnet werden, um eine Antwort auf unsere Frage zu erhalten.

Hier zeigt sich auch das Problem mit solchen Ansätzen, welches gleichzeitig auch die Antwort auf die Frage ist, wieso diese Form von künstlicher Intelligenz anders funktioniert als menschliche Intelligenz. Neuronale Netze wurden mit vielen Beispielen trainiert und damit die konkreten Zahlenverbindungen zwischen den Knoten festgelegt. Wird eine Ausgabe generiert, werden diese Knoten aktiviert und deren Zahlenwerte verrechnet. Wenn Menschen Entscheidungen treffen und Fragen beantworten, tun sie das allerdings nicht nur mit Zahlen und Berechnungen, sondern sie haben verschiedene gelernte Informationen im Kopf – vorherige Erfahrungen, auf denen sie aufbauen können, sie wissen von abstrakten Konzepten und Gruppierungen (Montag ist ein Arbeitstag, aber nicht jeder Arbeitstag ist ein Montag) und all das führt zu einem Bauchgefühl und Intuition. Etwas kann sich richtig oder falsch anfühlen und wir können konkret sagen, welche Informationen wir für die Entscheidung verwendet haben. Ein neuronales Netz hat diese Art des Abgleichs und der Erfahrung nicht; es sagt lediglich Zahlen vorher.

Large Language Models (LLMs) wie ChatGPT sind eine Weiterentwicklung solcher neuronalen Netze. Sie wurden mit unvorstellbar großen Datenmengen trainiert, können Text in Zahlen übersetzen, diese Zahlen verarbeiten und dann wiederum zurück in Text übersetzen. Wenn wir also eine Frage stellen („Was ist die Hauptstadt von Österreich?") wird nicht nach einer korrekten Antwort im eigenen Wissen gesucht, wie Menschen das tun würden. Stattdessen wird das statistisch am wahrscheinlichsten auftretende nächste Wort vorhergesagt („Wien"). Wird die Hauptstadt morgen zu Linz geändert, erfahren wir als Mensch diese neue Information und können unsere gespeicherten Informationen entsprechend anpassen. Ein LLM hingegen baut auf vielen Daten auf, in denen die Wörter „Hauptstadt", „Österreich" und „Wien" sehr häufig verknüpft sind und wird dementsprechend weiterhin das statistisch wahrscheinlichste Ergebnis ausgeben: Wien. Das LLM kann also Text ausgeben, ohne diesen Text selbst wirklich zu verstehen. Somit lässt sich erklären, weshalb Antworten falsch sein können: ChatGPT und andere LLMs besitzen keine große Wissensdatenbank mit allen Fragen, die die Menschheit je gestellt hat. Sie sagen lediglich das wahrscheinlichste nächste Wort vorher. Immer und immer wieder. Bis daraus ein ganzer Text wird. Dass das überhaupt funktioniert, ist an sich schon beeindruckend. Dass es damit auch noch oft richtig liegt, ist noch viel spannender. Trotzdem darf man nicht vergessen, dass diese Wortvorhersagen eben nicht mehr sind als genau das, Wortvorhersagen. Die Vorhersagen werden damit oft richtig und plausibel klingen, müssen aber nicht wahr sein.

Probleme bei der Verwendung von Large Language Models

Wie beschrieben, lernen LLMs im Training Sprachmuster und Begriffe, um das nächste Wort bestmöglich vorherzusagen. Zusätzlich werden viele Modelle aber auch noch nachtrainiert, damit sie Anweisungen besser befolgen und noch

hilfreicher wirken, zum Beispiel, indem zwei Antworten präsentiert werden und der Nutzer entscheiden soll, welche besser ist (4). Das erhöht zwar die Nutzbarkeit, aber möglicherweise auch die Gefahr, dass Antworten sehr sicher klingen, obwohl Unsicherheit angebrachter wäre. Schließlich wollen wir auf unsere Fragen nur selten „das weiß ich nicht" als Antwort erhalten.

Das alles führt dann zu sogenannten „Halluzinationen": Plausible, aber erfundene und nicht verifizierbare Angaben. Das können ausgedachte Leitlinien sein, Dosierungsempfehlungen, Studien oder auch Zitate, die in dieser Form überhaupt nicht existieren. Leider ist das auch kein Problem, das sich in zukünftigen Versionen beheben lässt, da diese Form von KI-Tools eben explizit zur Wortvorhersage entwickelt wurde und nicht als Entscheidungshilfe oder Wissensdatenbank. Für KI-Tools, die – ähnlich wie Menschen – Wissen und abstrakte Konzepte speichern und abrufen können, braucht es eine andere Form von künstlicher Intelligenz. Daran wird zwar bereits gearbeitet, der große Durchbruch lässt aber noch etwas auf sich warten.

Halluzinationen lassen sich also nicht so einfach beheben. Das ist besonders ungünstig, da geschätzt wird, dass je nach Modell generell 3-9% der Antworten „halluziniert" sind, in manchen Fragen sogar bis zu 82% (5,6), was zu falschen und potenziell gefährlichen Empfehlungen führen kann. Neben diesen falschen Empfehlungen, deren Konsequenz offensichtlich ist, gibt es aber auch noch die sogenannten Auslassungsfehler. Hierbei werden wichtige Informationen überhaupt nicht genannt (wie zum Beispiel, dass Bromsalz schnell zu Vergiftungen führen kann und nicht einfach als Salzersatz konsumiert werden sollte). Diese scheinen den Großteil von gefährlichen Fehlern auszumachen: In einer aktuellen Studie gab ChatGPT-5 in 17 von 100 Fällen potenziell gefährliche Empfehlungen, Google Gemini in 20 von 100 Fällen (7). Dabei waren 77% aller Fehler Auslassungsfehler (7). Denken wir zurück an den Anfang des Artikels und die Metakognition: Wenn wir keine Experten in einem Fachgebiet sind, wissen wir häufig gar nicht, was wir nicht wissen. Das macht es umso schwieriger, zu erkennen, wann ChatGPT uns wichtige Informationen vorenthält.

Natürlich können wir Rückfragen an ChatGPT stellen. Hier kommt aber ein weiteres Problem, das „sykophantische" Verhalten, ins Spiel. Viele LLMs wurden nämlich entwickelt, um für Nutzer*innen hilfreich zu sein. Deshalb neigen sie dazu, Zustimmung zu signalisieren oder sich einer bestimmten Erwartung anzupassen. Wenn ich also schreibe, dass 38 Grad Fieber doch nicht so schlimm sind und es bestimmt nur ein Infekt ist, der morgen weg ist, wird mir ChatGPT zustimmen: „Stimmt, nicht so schlimm". Wenn ich allerdings schreibe, dass ich mir große Sorgen mache, weil 38 Grad Fieber so hoch sind, ich mich fürchterlich krank fühle und am liebsten in die Notaufnahme gehen will, wird mir ChatGPT auch zustimmen: „Stimmt, das klingt gefährlich. Besser in die Notaufnahme". Diese Selbstbestätigung bei gleichzeitig möglicherweise fehlendem Wissen macht die Nutzung problematisch. Wir fühlen uns bestätigt und immer sicherer in unserer Meinung, egal, ob diese richtig oder falsch ist.

Das letzte Problem ist die Websuche. Manche Modelle können nun selbst das Internet durchsuchen, Webseiten lesen und Texte von diesen Webseiten einbeziehen. Das scheint ein Schritt in die richtige Richtung zu sein, bedeutet aber nicht, dass die Empfehlungen nun auch evidenzbasiert sind. Manche Webseiten sperren explizit den Zugang für LLMs, und es findet zudem keinerlei Qualitätskontrolle bei dieser Internetsuche statt. Eine Webseite mit evidenzbasierten Leitlinien hat also genau denselben Einfluss auf die Empfehlung wie große Werbeversprechen auf der Webseite eines Medizinprodukteherstellers – vorausgesetzt die Leitlinien werden überhaupt bei der Suche gefunden.

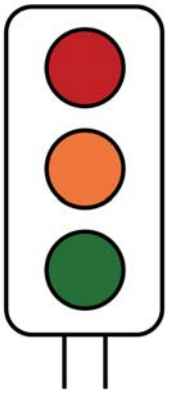
Mit all diesen Problemen ist es also weniger relevant, ob ein Modell eine niedrigere Halluzinationsrate hat, weniger „sykophantisch" ist oder das Internet durchsuchen kann. Viel wichtiger ist es, ob Nutzer*innen Informationen richtig einordnen und prüfen können.

Risiko als Leitprinzip: Welche Fragen eignen sich?

Wir können nicht Expert*innen in allen Fachgebieten sein, das ist klar – auch wenn es sich mit der Hilfe von ChatGPT manchmal so anfühlt. Worin wir aber Experte sein können, ist, die richtigen Fragen zu stellen. Gerade für die Hebammenpraxis und bei medizinischen Fragen generell sollte Risiko das Leitprinzip sein. Risiko heißt hier nicht nur die Wahrscheinlichkeit, dass etwas falsch ist, son-

dern vor allem: Welche Folgen hätte es, wenn ich falsch informiert wäre? Die Weltgesundheitsorganisation, die Europäische Union, und viele weitere Organisationen bewerten den Einsatz von KI-Systemen besonders anhand ihres Schadenspotenzials (8,9). Das gleiche Prinzip können wir auch auf unsere eigene Nutzung übertragen und bewerten, wie hoch der Schaden sein kann (Abb. 2).

Ein hohes Schadenspotenzial liegt vor, wenn Fehler akut gefährlich sein können, zu verzögerter Versorgung, falscher Behandlung, falscher (Selbst-)Medikation oder riskantem Verhalten führen können.



Hohes Risiko: Fehler können akut gefährlich sein, zu verzögerter Versorgung, falscher Behandlung oder riskantem Verhalten führen

Mittleres Risiko: Fehler können Handlungen beeinflussen, sind aber mit dem Prüfen von Quellen korrigierbar

Geringes Risiko: Fehler haben keine direkten negativen Auswirkungen oder sind leicht erkennbar und korrigierbar

Abb. 2. Risiko-Ampel für den Umgang mit KI-Chatbots.

Das ist in manchen Anwendungsfällen offensichtlich (Einschätzung, ob bestimmte Symptome einen Notfall darstellen), in anderen aber vielleicht nicht direkt klar (Alternativen für salzarme Ernährung).
Mögliche Anliegen mit einem hohen Risiko könnten sein: Soll ich in die Notaufnahme oder kann ich abwarten? Welche Diagnose erklärt meine Symptome? Wie interpretiere ich diesen Laborwert/Befund? Welche Dosis eines Medikaments brauche ich und welche Wechselwirkungen könnten entstehen? In all diesen Fällen sollten LLMs keine Zweitmeinung und erst recht keine Erstmeinung sein. Wenn überhaupt sollten LLMs hier nur genutzt werden, um sich auf die Kommunikation vorzubereiten, zum Beispiel um Symptome in Gruppen oder in Stichpunkten zu strukturieren. Liegt Expertise in dem Bereich vor (zum Beispiel bei Ärzt*innen, die nach möglichen Differentialdiagnosen fragen), können LLMs hilfreiche Hinweise geben. Wichtig ist aber eben, dass man ausreichend eigenes Wissen und Erfahrung hat, um diese Informationen richtig einzuordnen.

Ein mittleres Schadenspotenzial liegt vor, wenn Fehler zwar Handlungen (auf eine ungefährliche Art!) beeinflussen können, diese aber mit dem Prüfen von Quellen korrigierbar sind. So können beispielsweise Fachbegriffe oder typische Abläufe erklärt werden, verschiedene Optionen gegenübergestellt werden, oder konkrete Dokumente wie Leitlinien, Patienteninformationen oder Merkblätter zusammengefasst werden. Dennoch sollten auch hier Empfehlungen oder Informationen nicht unkritisch übernommen werden. Vielmehr sollten LLMs in Anwendungsfällen mit mittlerem Schadenspotenzial dazu genutzt werden, die Informationssuche zu starten. Fachbegriffe können im Anschluss selbst nochmal in zuverlässigen Quellen recherchiert und geprüft werden, und Begriffe aus Zusammenfassungen können die Suche in der vollständigen Leitlinie erleichtern.

Ein niedriges Schadenspotenzial liegt vor, wenn Fehler keine direkten negativen Auswirkungen haben oder leicht erkennbar und korrigierbar sind. Hier können beispielsweise Texte in einfachere Sprache umgeschrieben werden, komplexe Sachverhalte einfacher erklärt werden, der Schreibstil angepasst werden oder Texte gekürzt werden. Das ist besonders in hoch strukturierten Aufgaben nützlich: Briefe können mit LLMs geschrieben werden, Checklisten können erstellt werden, Zusammenfassungen von Dokumenten

VORGEHEN beim lateralen Lesen

1. Aussage isolieren:
Was genau wird vom LLM behauptet?
2. Mehrere unabhängige Quellen suchen:
z.B. Leitlinie/Fachgesellschaft und öffentliche Gesundheitsinstitution
3. Zuverlässigkeit der Quellen prüfen:
Ist der Bereitsteller dieser Informationen vertrauenswürdig?
4. Konsens prüfen:
Stimmen Aussagen überein? Widerspricht irgendetwas der LLM-Aussage oder gibt es weiteren Kontext?
5. Bei Unsicherheit Expert*innen fragen:
Bei Widerspruch, Unsicherheit oder hohem Risiko Experten befragen.

können geschrieben werden. Um sich noch mehr Zeit zu sparen, kann man auch mit Hilfe der Diktierfunktion einen langen Text aufnehmen und durch ein LLM automatisch strukturieren und/oder zusammenfassen lassen. All das sind Aufgaben, bei denen sich die Ergebnisse leicht überprüfen lassen und die kein hohes Schadenspotenzial besitzen.
Besonders bei Anwendungsfällen mit geringem oder mittlerem Risiko kann das Prüfen von Quellen und Informationen mit einer Taktik erleichtert werden, die sich laterales Lesen nennt (10) (Infobox 1). Ursprünglich als Maßnahme gegen Falschinformationen entwickelt, lässt sich diese auch auf Informationen von LLMs übertragen. Die Idee ist, dass eine Information nochmal durch eine Recherche in verschiedenen Quellen kritisch geprüft wird und dabei auch besonders die Zuverlässigkeit der Quellen bewertet wird.

Ausblick & Fazit

LLMs werden rasant weiterentwickelt, um diese verschiedenen Probleme zu adressieren: Neuere ChatGPT-Versionen sind weniger „sykophantisch", Websuchen zeigen die direkten Links als Quelle an, und auch neue Methoden zum Einbezug von externen Dokumenten („retrieval-augmented generation / RAG") werden zunehmend entwickelt und eingesetzt. Die Entwicklung geht auch in Richtung agentischer Systeme, das heißt, KI, die nicht nur antwortet, sondern Aufgaben ausführt – sucht, vergleicht, zusammenfasst, Termine plant oder Dokumente verarbeitet.
Dennoch bleibt die zentrale Frage: Wer prüft Qualität, Kontext und Risiko? KI wird im Alltag bleiben, egal ob Sie persönlich gegen oder für die Nutzung sind. Die Aufgabe sollte auch gar nicht sein, diese Entwicklung aufzuhalten, sondern viel mehr, Anwender*innen dabei zu unterstützen, KI so zu nutzen, dass ihre Gesundheitskompetenz gestärkt und Risiken reduziert werden. Dabei können die folgenden Merksätze helfen:

(1) Plausibel ist nicht gleich richtig.
(2) Je höher das Risiko, desto weniger geeignet sind KI-Chatbots.
(3) Bei Unsicherheit ist professionelle Abklärung wichtig.
(4) KI-Ausgaben sollten immer nochmal mit einer Suche in zuverlässigen Quellen überprüft werden.

Mit diesen Informationen kann KI im Alltag hoffentlich sinnvoller genutzt werden – und gleichzeitig dafür sorgen, dass ein Kochrezept vielleicht kürzer wird, das Speisesalz aber nicht durch Bromsalz ersetzt wird.

Literaturverzeichnis

1. Eichenberger A, Thielke S, Van Buskirk A. A Case of Bromism Influenced by Use of Artificial Intelligence. AIM Clinical Cases. 2025 Aug;4(8):e241260.
2. Kopka M, Wang SM, Kunz S, Schmid C, Feufel MA. Technology-supported self-triage decision making. npj Health Syst. 2025 Jan 25;2(1):1–11.
3. Takita H, Kabata D, Walston SL, Tatekawa H, Saito K, Tsujimoto Y, et al. A systematic review and meta-analysis of diagnostic performance comparison between generative AI and physicians. npj Digit Med. 2025 Mar 22;8(1):175.
4. Ouyang L, Wu J, Jiang X, Almeida D, Wainwright CL, Mishkin P, et al. Training language models to follow instructions with human feedback. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc.; 2022. p. 27730–44. (NIPS '22).
5. Hughes S, Bae M, Li M. Vectara Hallucination Leaderboard [Internet]. 2023 [zitiert 07. Januar 2026]. Verfügbar unter: https://github.com/vectara/hallucination-leaderboard
6. Omar M, Sorin V, Collins JD, Reich D, Freeman R, Gavin N, et al. Multi-model assurance analysis showing large language models are highly vulnerable to adversarial hallucination attacks during clinical decision support. Commun Med. 2025 Aug 2;5(1):330.
7. Wu D, Haredasht FN, Maharaj SK, Jain P, Tran J, Gwiazdon M, et al. First, do NOHARM: towards clinically safe large language models [Internet]. arXiv; 2025 [zitiert 07. Januar 2026]. Verfügbar unter: http://arxiv.org/abs/2512.01241
8. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance) [Internet]. Jun 13, 2024. Verfügbar unter: http://data.europa.eu/eli/reg/2024/1689/oj
9. Ethics and governance of artificial intelligence for health: Guidance on large multi-modal models [Internet]. [zitiert 11. Januar 2026]. Verfügbar unter: https://www.who.int/publications/i/item/9789240084759
10. McGrew S. Teaching lateral reading: Interventions to help people read like fact checkers. Current Opinion in Psychology. 2024 Feb 1;55:101737.

Dr. Dr. MARVIN KOPKA, MPH

ist wissenschaftlicher Mitarbeiter am Fachgebiet Arbeitswissenschaft der Technischen Universität Berlin, wo er zu künstlicher Intelligenz in Medizin und Public Health forscht. Ein besonderer Schwerpunkt liegt dabei auf Entscheidungen und Verhalten und wie KI diese beeinflusst.
Kontakt: mail@marvinkopka.com