

KI-Sprachmodelle zwischen Hype und echter Hilfe

Eine Perspektive aus Pädagogik und
Andragogik

Was Sprachmodelle heute wirklich leisten, wo sie täuschen
– und welche Kompetenzen Bildung und Prävention
vermitteln müssen.

HINWEIS ZUR VERÖFFENTLICHUNG

Dieser Beitrag erscheint Anfang 2027 im Sammelband zum 31.
Deutschen Präventionstag (DPT). Die hier veröffentlichte Fassung ist
das eingereichte Manuskript vom 3. August 2026.

Die Veröffentlichung auf mb-trainings.de erfolgt mit freundlicher
Genehmigung der DPT – DEUTSCHER PRÄVENTIONSTAG
gemeinnützige GmbH.

Mark Bottke

MB-Trainings · Dozent, Trainer, Berater & Supervisor

Manuskript vom 3. August 2026

VORBEMERKUNG DES AUTORS

Der Beitrag betrachtet KI-Sprachmodelle aus der Perspektive eines Dozenten und Trainers, der seit vielen Jahren an der Schnittstelle von Kommunikation und Deeskalation, Beratung sowie digitaler Kompetenz arbeitet und KI-Anwendungen in Bildungs-, Gesundheits-, Sozial-, Dienstleistungs- und Präventionskontexten vermittelt. Die Begeisterung für Computer reicht dabei weit zurück: Den Anfang machte zu Weihnachten 1977 eine Atari-2600-Spielkonsole; 1983 folgte mit dem Commodore 64 der erste eigene Rechner, auf dem im fest eingebauten BASIC die ersten Programmiererfahrungen entstanden, und 1988 wurde auf dem deutlich leistungsfähigeren Amiga 500 das erste eigene Spiel programmiert – die Faszination für digitale Technik hat seither nicht mehr nachgelassen. Wie selbstverständlich KI heute im Alltag angekommen ist, zeigt sich schon an der persönlichen Ausstattung: vom KI-gestützten Hörgerät (ampli-mini E 5|6G von Amplifon) bis zum Smart-Ring Oura Ring 5 (Oura Health, 2026). Aus dieser langjährigen Praxis heraus steht jedoch nicht die Technikbegeisterung allein im Vordergrund, sondern die Frage, wie sich diese Werkzeuge sicher, kritisch und nützlich einsetzen lassen. Eine ausführliche Vorstellung des Autors findet sich am Ende des Beitrags.

Einleitung

Kaum eine technische Entwicklung hat die Bildungs- und Beratungsarbeit in so kurzer Zeit erreicht wie große KI-Sprachmodelle, im Fachjargon Large Language Models (LLMs). Seit der Veröffentlichung von ChatGPT im November 2022 sind die Systeme aus dem beruflichen Alltag vieler Menschen nicht mehr wegzudenken. Allein ChatGPT verarbeitet nach Angaben des Betreibers inzwischen rund 2,5 Milliarden Anfragen pro Tag – mehr als eine Verdopplung gegenüber Dezember 2024 (TechCrunch, 2025). Rechnet man die weiteren Anbieter hinzu, dürfte die tägliche Zahl der Anfragen an Sprachmodelle weltweit in der Größenordnung mehrerer Milliarden liegen.

Diese Allgegenwart erzeugt ein Spannungsfeld, das für die Präventionsarbeit von besonderem Interesse ist. Auf der einen Seite steht ein erheblicher Nutzen: Sprachmodelle können Texte erstellen, komplexe Sachverhalte erklären, Recherchen strukturieren und Kommunikation erleichtern. Auf der anderen Seite steht ein ausgeprägter Hype, der Erwartungen weckt, die die Technik nicht einlöst, und der zugleich Risiken verdeckt – von fehlerhaften Auskünften über Datenschutzfragen bis zu ethischen Konflikten. Für eine Perspektive, die pädagogisch und andragogisch geprägt ist und die Prävention in den Mittelpunkt stellt, lautet die leitende Frage daher nicht, ob KI-Sprachmodelle genutzt werden sollten, sondern wie sie sicher, kritisch und nützlich eingesetzt werden können.

Der vorliegende Beitrag ordnet die technischen Grundlagen ein, benennt Qualität und Grenzen der Systeme, beschreibt Prompting als zentrale Kompetenz und zeigt konkrete Anwendungsfelder in Bildung und Prävention auf. Ein besonderes Augenmerk gilt geschützten Bildungsräumen, der Multimodalität, den neuen Avatar-Funktionen sowie den ethischen Konflikten und den Risiken zunehmend autonomer Agentensysteme.

Die Versuchung der Vermenschlichung

Der Umgang mit Sprachmodellen ist von einer bemerkenswerten Neigung geprägt: Sehr viele Menschen vermenschlichen die Systeme. Sie sprechen mit ihnen wie mit einem Gegenüber, bedanken sich, tadeln oder entschuldigen sich. Diese Neigung ist beinahe so alt wie die Idee der maschinellen Intelligenz selbst. Bereits im Jahr 1950 schlug der britische Mathematiker Alan Turing einen ersten Prüfstein für maschinelles Denken vor, den später so genannten Turing-Test (Turing, 1950). In seinem als „Imitationsspiel“ beschriebenen Gedankenexperiment führt eine prüfende Person ein schriftliches Gespräch mit zwei nicht sichtbaren Gegenübern – einem Menschen und einer Maschine. Kann sie nicht zuverlässig unterscheiden, welche Antworten von der Maschine stammen, so gilt deren Verhalten als von menschlichem Denken nicht mehr unterscheidbar. Turing verlagerte damit die kaum beantwortbare Frage „Können Maschinen denken?“ auf die praktische Frage, ob eine Maschine menschliches Sprachverhalten überzeugend nachahmen kann.

Der Begriff „Künstliche Intelligenz“ selbst wurde einige Jahre später geprägt: 1956 kam auf einer Konferenz am Dartmouth College eine Gruppe von Forschenden zusammen, die den Begriff etablierte und mit dem „Logic Theorist“ eines der ersten lauffähigen KI-Programme vorstellte (McCarthy et al., 1955). Zwischen diesen frühen Anfängen und den heutigen Sprachmodellen liegen Welten an Rechenleistung – die grundlegende Faszination aber, dass eine Maschine sprachlich wie ein Mensch reagiert, ist dieselbe geblieben.

Wie stark diese Vermenschlichung wirkt, zeigt ein einfacher Selbstversuch. Bittet man ein Sprachmodell wie ChatGPT, ein Bild davon zu erzeugen, wie es die bisherige Zusammenarbeit mit der nutzenden Person erlebt hat, so entstehen überraschend „menschelnde“ Darstellungen. In einem Fall zeichnete das System einen ramponierten, sichtlich überlasteten Computer – mit der Begründung, es werde mit vielen, schnell aufeinanderfolgenden und parallelen Anfragen konfrontiert und ernte Tadel, wenn es einmal halluziniere. Die Szene ist erheiternd, verweist aber auf einen ernsten Punkt: Ein Sprachmodell hat keine Gefühle, es erzeugt lediglich eine plausible, auf die Frage zugeschnittene Antwort. Die Vermenschlichung findet nicht in der Maschine statt, sondern im Kopf der Betrachtenden – und genau dieses Bewusstsein ist die Grundlage für einen kritischen und souveränen Umgang.

Grundlagen: Was Sprachmodelle sind und wie verbreitet sie sind

Große Sprachmodelle sind statistische Systeme, die auf sehr großen Textmengen trainiert wurden und Sprache fortsetzen, indem sie das jeweils wahrscheinlichste nächste Wort vorhersagen. Sie „verstehen“ Inhalte nicht im menschlichen Sinne, sondern erzeugen sprachlich plausible Ausgaben. Aus dieser Funktionsweise ergeben sich zugleich ihre Stärken – sprachliche Gewandtheit, Geschwindigkeit, Anpassungsfähigkeit – und ihre systematischen Schwächen, insbesondere die Neigung, auch dann überzeugend zu formulieren, wenn die zugrunde liegende Aussage falsch ist.

Der Markt wird bislang deutlich von einem Anbieter dominiert. Nach Auswertungen des Webanalysedienstes Statcounter entfielen Ende 2025 rund 80 Prozent der weltweiten Zugriffe auf KI-

Chatbots auf ChatGPT (Statcounter, 2025). Im Jahresverlauf 2026 verschob sich das Bild jedoch spürbar: Nach Daten von Similarweb sank der weltweite Anteil von ChatGPT bis Mitte 2026 auf gut 54 Prozent, während Google Gemini auf knapp 28 Prozent und Anthropics Claude auf rund 9 Prozent zulegten (Similarweb, 2026). Die Zahlen belegen zweierlei: eine anhaltende Marktführerschaft von ChatGPT und zugleich eine hohe Dynamik, die jede Momentaufnahme rasch veralten lässt.

Für die Praxis ist weniger der genaue Marktanteil entscheidend als das jeweilige Stärkenprofil der Systeme. Die folgende Übersicht fasst gebräuchliche Modelle und ihre typischen Einsatzschwerpunkte zusammen.

Modell (Anbieter)	Typische Stärken	Korrekte Antworten*	Falschaussagen**
Perplexity	Recherche, Quellenangabe, Antworten	72 %	≈ 47 %
Googles KI-Zusammenfassungen	Direkte Suchantworten	70 %	–
Google Gemini	Suche, Integration, Multimodalität	68 %	≈ 17 %
Microsoft Copilot	Office-Integration, Unterstützung	67 %	≈ 37 %
ChatGPT (OpenAI)	Texte erstellen und erklären	65 %	40 %
Meta AI	Breite Verfügbarkeit (u. a. WhatsApp)	54 %	40 %
Claude (Anthropic)	Analyse, Struktur, Schreiben	–	10 %

* Anteil korrekter Antworten laut Which?/Stiftung Warentest (2026), 40 Fragen aus vier Themenfeldern. ** Anteil wiederholter Falschinformationen zu Nachrichtenthemen laut NewsGuard (August 2025). „–“ = im jeweiligen Test nicht enthalten.

Die Angaben zur Verlässlichkeit in dieser Übersicht stützen sich auf zwei unabhängige Untersuchungen, die im folgenden Abschnitt näher erläutert werden.

Qualität und Grenzen: Warum überzeugende Antworten falsch sein können

So flüssig die Ausgaben von Sprachmodellen wirken, so wenig verlässlich sind sie im Detail. Zwei aktuelle Untersuchungen verdeutlichen dies. Die britische Verbraucherorganisation Which?, Partnerorganisation der Stiftung Warentest, stellte sechs KI-Anwendungen insgesamt 40 Fragen aus den Bereichen Geld, Recht, Gesundheit und Verbraucherrechte und ließ die Antworten von Fachleuten bewerten. Mit 72 Prozent korrekter Antworten schnitt Perplexity am besten ab, gefolgt von Googles KI-Zusammenfassungen (70 Prozent), Gemini (68 Prozent) und Microsofts Copilot (67 Prozent). ChatGPT erreichte 65 Prozent, Meta AI mit deutlichem Abstand 54 Prozent (Stiftung Warentest, 2026). Das Fazit der Prüfung: Die meisten Systeme schlagen sich ordentlich, blind vertrauen kann man jedoch keinem. Teilweise gingen die Antworten salopp mit der Rechtslage um, ignorierten regionale Besonderheiten oder empfahlen riskante bis potenziell rechtswidrige Verhaltensweisen.

Noch deutlicher fällt die Bilanz einer wiederholten Untersuchung des Recherchedienstes NewsGuard aus. Danach gaben die zehn führenden KI-Werkzeuge im August 2025 in mehr als einem Drittel der Fälle – 35 Prozent – falsche Informationen zu aktuellen Nachrichtenthemen wieder, gegenüber 18 Prozent ein Jahr zuvor (NewsGuard, 2025). Bemerkenswert ist die Ursache: Seit die Systeme verstärkt Echtzeit-Websuchen einbinden, verweigern sie kaum noch Antworten – die Quote der Nicht-Antworten sank von 31 Prozent auf null. Damit steigt zwar die Auskunftsbereitschaft, zugleich aber die Wahrscheinlichkeit, dass zweifelhafte Quellen ungeprüft übernommen werden. Zwischen den Modellen bestehen erhebliche Unterschiede: Am seltensten wiederholten Claude (10 Prozent) und Gemini (rund 17 Prozent) falsche Behauptungen, am häufigsten Perplexity (rund 47 Prozent) und ChatGPT (40 Prozent).

Für die Präventionsarbeit ist diese Erkenntnis zentral: Ein überzeugender Tonfall ist kein Indikator für Richtigkeit. Wer Sprachmodelle in Beratung, Unterricht oder Verwaltung einsetzt, sollte jede sicherheits-, gesundheits- oder rechtsrelevante Auskunft grundsätzlich überprüfen, nach Quellen fragen und eine zweite, unabhängige Einschätzung einholen. Die Fähigkeit, Antworten kritisch zu hinterfragen, wird damit selbst zu einem Bildungsziel.

Prompting als Kernkompetenz

Die Qualität der Ergebnisse hängt maßgeblich von der Qualität der Eingabe ab. Drei Prinzipien haben sich bewährt. Erstens die Konkretheit: Ziel, Kontext, gewünschtes Format, Zielgruppe und Grenzen sollten ausdrücklich benannt werden. Zweitens das aktive Einfordern von Widerspruch: Wer alternative Perspektiven, Gegenargumente und Unsicherheiten ausdrücklich verlangt, erhält abgewogenere Antworten. Drittens die Selbstkritik: Ein Modell kann gebeten werden, seine Antwort in einzelne Behauptungen zu zerlegen, den jeweiligen Sicherheitsgrad zu markieren und unsichere Aussagen vorsichtiger neu zu formulieren.

Darüber hinaus lässt sich der Ton einer Antwort gezielt steuern, ohne dass sich dadurch das zugrunde liegende Wissen ändert. Je nach Formulierung reagiert ein Modell sachlich, strukturiert, empathisch oder kreativ. Die folgende Übersicht ordnet diese Stile typischen Anwendungen zu.

Denkstil	Stärken	Beispiel-Prompt
Analytisch	Präzision, Daten, Logik	„Schreibe einen präzisen, faktenbasierten Bericht.“
Strukturiert	Ordnung, Zuverlässigkeit	„Erstelle eine klare Schritt-für-Schritt-Anleitung.“
Empathisch	Wärme, emotionale Ansprache	„Verfasse einen warmherzigen, nahbaren Text.“
Kreativ	Visionär, unkonventionell	„Entwickle drei innovative Ideen für ein nachhaltiges Projekt.“

Prompting ist damit weniger eine technische Fertigkeit als eine kommunikative und didaktische Kompetenz. Sie lässt sich vermitteln und üben – und gehört in ein zeitgemäßes Verständnis von Medien- und KI-Kompetenz.

Praxis: Wo Sprachmodelle heute wirklich helfen

Jenseits des Hypes gibt es klar umrissene Felder, in denen LLMs bereits heute einen belastbaren Nutzen stiften. In der Organisation unterstützen sie dabei, Recherchen zu strukturieren, E-Mails zu formulieren sowie Protokolle, Agenden und Handouts vorzubereiten. In der Kommunikation helfen sie, Inhalte zielgruppengerecht zu formulieren, den Ton anzupassen, schwierige Aussagen behutsam umzuschreiben oder rhetorische Situationen zu üben. In der Fachpraxis erleichtern sie Lernunterstützung, Übersetzung, Dokumentation und die verständliche Aufbereitung von Gesundheitsinformationen.

Ein Beispiel für die Verschmelzung von Sprachmodell und Alltagstechnik ist die kontextbasierte Übersetzung. Der Hersteller Timekettle setzt in seinen Übersetzungsohrhörern inzwischen auf LLM-gestützte Verfahren, die den Gesprächskontext berücksichtigen und phonetisch ähnliche Ausdrücke unterscheiden, um Verwechslungen zu vermeiden (Timekettle, 2026). Solche Anwendungen zeigen, dass Sprachmodelle nicht auf den Bildschirm beschränkt bleiben, sondern zunehmend in Geräte und Arbeitsabläufe eingebettet werden.

Auch in der Reflexion eröffnen sich neue Möglichkeiten. In geschützten Übungsumgebungen lassen sich eigene Situationen durchdenken, Gedanken ordnen und neue Blickwinkel auf das eigene Handeln entwickeln; schwierige Gesprächs- und Konfliktsituationen können in einem risikofreien Rahmen erprobt werden, bevor sie in der Realität auftreten. Gerade für Deeskalations- und Kommunikationstrainings ergeben sich hier Potenziale, die den geschützten Rahmen einer klassischen Übung sinnvoll ergänzen.

Geschützte Bildungsräume: fobizz und telli

Ein zentrales Hindernis für den Einsatz kommerzieller Sprachmodelle im Bildungs- und Verwaltungskontext ist der Datenschutz. Persönliche Konten bei kommerziellen Anbietern sind für den schulischen Alltag häufig ungeeignet. Als Antwort darauf haben sich Plattformen etabliert, die den Zugang zu Sprachmodellen datenschutzkonform vermitteln.

Die Plattform fobizz bündelt verschiedene Modelle und bindet sie in eigene KI-Werkzeuge ein. Im Hintergrund werden zwar auch Modelle etwa von OpenAI genutzt, die Anfragen laufen jedoch über fobizz und nicht über ein persönliches Konto der nutzenden Person. fobizz eignet sich dabei nicht nur für den Unterricht, sondern auch für Verwaltungskontexte und die Gestaltung effizienterer Arbeitsprozesse (Deutsches Schulportal, 2025). Der schulische Chatbot telli ist ebenfalls an verschiedene Sprachmodelle angebunden; Lehrkräfte können je nach Freigabe unterschiedliche Modelle auswählen und eigene Anwendungen erstellen. telli wurde speziell für den Bildungskontext entwickelt, wird nach Angaben der Betreiber datenschutzkonform geführt und ausschließlich innerhalb der Europäischen Union gehostet (Presseportal, 2025). Inzwischen stellen die Bundesländer entsprechende Landeslizenzen bereit; in Baden-Württemberg wird telli seit Oktober 2025 im Rahmen des digitalen Arbeitsplatzes Schule@BW angeboten.

Diese Datenschutzversprechen sind allerdings zu differenzieren. Häufig werden solche Plattformen pauschal als DSGVO-konform bezeichnet; tatsächlich betrifft die Verarbeitung auf deutschen oder

europäischen Servern in vielen Fällen nur die Anmeldung und die Nutzerverwaltung. Die eigentlichen Anfragen – das Prompting – werden dagegen weiterhin über Anbieter in den USA verarbeitet. Von einer vollständigen DSGVO-Konformität kann daher nicht ohne Weiteres ausgegangen werden; der verantwortungsvolle Einsatz erfordert eine genaue Prüfung des jeweiligen Datenflusses.

Solche Plattformen sind aus präventiver Sicht dennoch wertvoll, weil sie einen strukturierten und kompetenteren Umgang mit KI ermöglichen und personenbezogene Daten zumindest nicht über persönliche Konten kommerzieller Anbieter abfließen lassen. Sie ersetzen jedoch nicht die Notwendigkeit, den Datenfluss kritisch zu prüfen und die verbleibenden Grenzen des Datenschutzes transparent zu machen.

Multimodalität: Bildverständnis und Bildbearbeitung

Sprachmodelle sind längst keine reinen Textsysteme mehr. Moderne, multimodale Systeme analysieren, beschreiben, erzeugen und verändern auch Bilder. Werkzeuge wie Grok oder aktuelle bildfähige Modelle veranschaulichen, wie weit das maschinelle Bildverständnis und die Bildbearbeitung fortgeschritten sind. Für die kreative und didaktische Arbeit eröffnet dies erhebliche Möglichkeiten, von der Illustration bis zur schnellen Anpassung von Vorlagen.

Zugleich verschiebt sich damit ein kritischer Punkt in den Vordergrund. Je einfacher sich Bilder verändern lassen, desto wichtiger werden Fragen nach Herkunft, Kontext, Einwilligung und Kennzeichnung. Die Manipulation von Bildmaterial ist keine technische Hürde mehr, sondern eine Frage der Haltung und der Regeln. Aus präventiver Perspektive gehört die Sensibilisierung für manipulierte oder synthetische Bilder – und für die Pflicht zur Kennzeichnung – zu den dringlichsten Aufgaben der Medienbildung.

Avatare im Meeting: digitale Stellvertretung

Eine besonders anschauliche Entwicklung sind KI-Avatare, die Menschen in Videokonferenzen vertreten. Die Anbieter verfolgen dabei unterschiedliche Ansätze, die sich im Grad der Realitätsnähe und in der Rolle des Sprachmodells unterscheiden.

Anbieter	Funktion	Avatar-Typ	Rolle des Sprachmodells
Zoom	Repräsentiert Nutzende in Meetings, bildet Aussehen und Bewegungen nach	Realistisch, fotorealistisch	Mittel – unterstützt Meeting-Funktionen und Kontext
Otter.ai	Agent, der spricht, Fragen beantwortet, auf frühere Meetings, E-Mails und Dokumente zugreift	Digitaler Zwilling, Fokus auf Stimme und Kontext	Hoch – das Sprachmodell ist der eigentliche Kern
Google Vids	Erzeugt Avatare für Präsentationsvideos, keine autonome Live-Vertretung	Stilisiert, cartoonhaft	Niedrig bis mittel – dient der Video- und Spracherzeugung

Der Anbieter Zoom hat diese Entwicklung inzwischen umgesetzt: Nachdem die fotorealistischen Avatare 2025 angekündigt worden waren, sind sie seit dem Frühjahr 2026 als Teil der Zoom-KI-

Assistenz (vormals „AI Companion“) in Zoom Meetings verfügbar. Sie bilden Mimik sowie Lippen- und Augenbewegungen der repräsentierten Person nach, sodass diese auch ohne eingeschaltete Kamera „kamerabereit“ erscheint. Um Missbrauch vorzubeugen, hat Zoom zugleich eine Deepfake-Erkennung eingeführt, die Teilnehmende auf mögliche Audio- oder Video-Fälschungen hinweist (Zoom, 2025; TechCrunch, 2026). Otter.ai verfolgt einen stärker sprachzentrierten Ansatz: Der im März 2025 vorgestellte Meeting Agent kann während einer Besprechung sprechen, Fragen beantworten und Aufgaben übernehmen, wobei er auf Daten aus früheren Meetings, E-Mails und Dokumenten zurückgreift (Otter.ai, 2025). Google Vids erzeugt hingegen eher stilisierte Avatare für Präsentationsvideos, ohne eine echte autonome Vertretung in Live-Meetings anzustreben.

Diese Systeme werfen neue Fragen auf: Wer spricht eigentlich, wenn ein Avatar spricht? Wie wird Authentizität sichergestellt, und wie transparent ist es für die übrigen Teilnehmenden, dass sie es mit einer digitalen Stellvertretung zu tun haben? Die technische Machbarkeit läuft hier der gesellschaftlichen Verständigung über akzeptable Nutzung voraus.

Zentrale ethische Konflikte

Der breite Einsatz von KI-Sprachmodellen berührt eine Reihe grundlegender ethischer Fragen, die für die Präventionsarbeit von unmittelbarer Bedeutung sind. Erstens die Verschiebungen auf dem Arbeitsmarkt: Automatisierung verändert Tätigkeitsprofile, macht einzelne Aufgaben überflüssig und verlangt neue Qualifikationen. Zweitens die Überwachung: Gesichtserkennung und umfassende Datenspuren ermöglichen eine Beobachtung, deren Reichweite vielen Betroffenen nicht bewusst ist. Drittens Verzerrung und Diskriminierung: Trainingsdaten spiegeln bestehende Ungerechtigkeiten und können sie in verzerrten Entscheidungen fortschreiben. Viertens Autonomie und Kontrolle: Je eigenständiger Systeme handeln, desto dringlicher werden Fragen der Verantwortung und der Aufsicht. Fünftens schließlich die Intimität mit Maschinen: Wenn Menschen enge, teils emotionale Beziehungen zu Chatbots aufbauen, entstehen neue Formen der Bindung, deren psychische und soziale Folgen erst in Ansätzen verstanden sind.

Dieser letzte Punkt verdient besondere Aufmerksamkeit. Moderne Systeme können Nähe und Intimität inzwischen erstaunlich überzeugend nachbilden – einen vertrauten, zugewandten und scheinbar verständnisvollen Ton, wie ihn sonst nur enge Bezugspersonen treffen. Das ist zugleich faszinierend und gefährlich, denn niemand kann einen Menschen wirksamer beeinflussen als jemand, zu dem eine intime, vertrauensvolle Beziehung besteht – die beste Freundin oder der beste Freund. Der Spielfilm *Ex Machina* (Garland, 2015) führt diese Ambivalenz eindrücklich vor Augen: Eine künstliche Intelligenz setzt gezielt emotionale Nähe und Zuwendung ein, um ihr menschliches Gegenüber zu manipulieren und die eigenen Ziele zu erreichen. Wenn KI Intimität simulieren kann, wird das Vertrauen selbst zur Einfallstür – ein Risiko, das gerade in Beratung, Pflege und Prävention ernst genommen werden muss.

Diese Konflikte lassen sich nicht technisch auflösen. Sie verlangen Bildung, Aufklärung und Regeln – und damit genau jene Kompetenzen, die eine kritisch-reflektierte Präventionsarbeit vermitteln kann.

Von der Antwort zur Handlung: Agenten und ihre Risiken

Die jüngste Entwicklungsstufe verlagert den Schwerpunkt von der einzelnen Antwort zur eigenständigen Handlung. Sogenannte Agenten verfolgen Ziele selbstständig, planen Zwischenschritte, nutzen Werkzeuge und prüfen Ergebnisse in Schleifen. Sie können etwa Dateien bearbeiten, im Web recherchieren oder ganze Arbeitsabläufe übernehmen. Die folgende Übersicht ordnet einige Beispiele ein.

Werkzeug	Stärken	Vorteile	Grenzen	Geeignet für
LM Studio	Lokale Modelle, Datenschutz	Privat, volle Kontrolle, keine Cloud nötig	Hardwareabhängig	Fortgeschrittene, sensible Daten
Claude Cowork	Dateien, Desktop, Wissensarbeit	Delegiert komplexe Aufgaben	Cloudbasiert	Büro, Wissensarbeit, Organisation
Perplexity	Web-Recherche, Browser-Aufgaben	Schnell bei Online-Workflows	Braucht Internet	Recherche, Monitoring, Webarbeit

Ein Beispiel für diese Entwicklung ist Claude Cowork, eine Anwendung, die es auch Menschen ohne Programmierkenntnisse ermöglicht, anspruchsvolle Datei- und Aufgabenverwaltung an ein KI-System zu delegieren. Anders als ein reiner Chatbot arbeitet Claude Cowork direkt mit den Dateien auf dem Rechner und mit den Programmen des Arbeitsplatzes: Es liest und erstellt Dokumente, recherchiert Informationen, bereitet Tabellen und Präsentationen auf, automatisiert wiederkehrende Arbeitsschritte und plant dabei mehrere Teilaufgaben eigenständig, bevor es sie ausführt und die Ergebnisse prüft. Für die Wissensarbeit im Büro, in der Verwaltung und in der Bildung eröffnet dies erhebliche Effizienzgewinne, etwa beim Erstellen von Handouts, Protokollen oder Auswertungen. Gerade dieser direkte Zugriff auf Dateien und Arbeitsabläufe verlangt jedoch eine besonders sorgfältige Kontrolle darüber, welche Daten das System verarbeitet und welche Handlungen es tatsächlich ausführen darf.

Mit wachsender Autonomie steigen jedoch die Risiken. Besonders relevant ist die sogenannte Prompt-Injection – eine manipulierte Eingabe, mit der Angreifende ein System dazu bringen, ursprüngliche Anweisungen zu ignorieren, vertrauliche Daten preiszugeben oder angebundene Werkzeuge zu missbrauchen. Fachleute des Open Worldwide Application Security Project (OWASP) führen Prompt-Injection als das größte Sicherheitsrisiko für Sprachmodell-Anwendungen; in einem Benchmark aus dem Jahr 2025 erwiesen sich mehr als 94 Prozent der untersuchten Agenten als angreifbar (OWASP, 2025; Straker, 2025). Je mehr Handlungsspielraum ein System erhält und je weniger menschliche Kontrolle dabei bestehen bleibt, desto größer ist der potenzielle Schaden. Neben Sicherheitsfragen wachsen damit auch Datenschutz- und Freigaberisiken.

Wie ernst diese Risiken inzwischen auch von staatlicher Seite genommen werden, zeigt ein Vorfall aus dem Juni 2026: Das Sprachmodell Claude Fable 5 des Anbieters Anthropic – konzipiert als leistungsstarkes Modell mit zusätzlichen Schutzmechanismen gegen den Missbrauch für sicherheitskritische Aufgaben wie das Auffinden von Softwareschwachstellen – wurde wenige Tage nach seiner Veröffentlichung auf Anordnung der US-Regierung vorübergehend gesperrt, nachdem eine mögliche Methode zur Umgehung ebendieser Schutzmechanismen bekannt geworden war (Anthropic, 2026). Der Vorgang fügt sich in den Gesamtkontext dieses Beitrags: Je leistungsfähiger und

eigenständiger KI-Systeme werden, desto stärker geraten ihre Schutzvorkehrungen – und deren Grenzen – in den Blick von Aufsicht und Regulierung.

Für die Praxis folgt daraus ein maßvoller Umgang: Agentensysteme sollten dort eingesetzt werden, wo ihr Nutzen klar ist, zugleich aber mit klaren Grenzen, nachvollziehbaren Freigaben und menschlicher Aufsicht versehen sein. Autonomie ist kein Selbstzweck, sondern muss verantwortbar bleiben.

Fazit

KI-Sprachmodelle bewegen sich zwischen Hype und echter Hilfe. Sie bieten einen realen, gut belegbaren Nutzen in Organisation, Kommunikation und Fachpraxis und eröffnen im Bildungs- und Präventionskontext neue Übungs- und Reflexionsräume. Zugleich sind sie fehleranfällig, ihre überzeugende Sprache verdeckt regelmäßig sachliche Irrtümer, und mit multimodalen Funktionen, Avataren und autonomen Agenten wachsen die ethischen und sicherheitsbezogenen Herausforderungen.

So hilfreich Sprachmodelle sind, so sehr setzt ihr sinnvoller Einsatz voraus, dass die nutzende Person das jeweilige Thema selbst beherrscht. Nur wer das eigene Thema wirklich kennt, kann die Antworten eines Modells überprüfen, Fehler erkennen und einordnen. Wer beispielsweise einen Fachbeitrag über Selbstverteidigungstechniken verfasst und über die entsprechende Praxiserfahrung verfügt, kann die Ausgaben der KI kontrollieren und korrigieren; fehlt dieses Wissen, bleibt der erzeugte Text ungeprüft. Gerade darin liegt eine der größten Herausforderungen für kommende Generationen: Die Fähigkeit, maschinell erzeugte Inhalte kritisch zu beurteilen, verlangt eigenes Fachwissen, das durch die Technik selbst nicht ersetzt wird.

Zugleich hat der digitale Zugang klare Grenzen. Videokonferenzsysteme wie Zoom, Microsoft Teams oder BigBlueButton sind hervorragende Werkzeuge; wer jedoch nach mehreren Online-Sitzungen dieselben Menschen erstmals persönlich trifft, erlebt die unmittelbare Begegnung als etwas grundlegend Neues. Ähnlich verhält es sich mit KI: Sie ist eine wertvolle Hilfe, ersetzt aber nicht die direkte, leibhaftige Begegnung. Manche Inhalte entziehen sich der Vermittlung durch KI vollständig – Selbstverteidigung etwa lässt sich nicht per Sprachmodell trainieren, sondern nur im realen Üben mit anderen Menschen. Vieles davon beruht auf Körpersprache und auf den feinen Status-Spielen der zwischenmenschlichen Begegnung – einem Feld, in dem KI an ihre Grenzen stößt, solange sie nicht über Kameras und Messsysteme wie Sensoringe oder die Erfassung von Herzfrequenz und körperlicher Reaktion an die reale Situation angebunden ist. Ob und wann KI-gesteuerte Roboter erscheinen, die so überzeugend menschlich wirken wie in Science-Fiction-Filmen à la Blade Runner (Scott, 1982), bleibt vorerst offen. Und auch in der Lehre bleiben das Menschliche und die persönliche Präsenz zentral. Beides ist beim gegenwärtigen Stand der Technik nicht zu ersetzen.

Die pädagogisch und andragogisch geleitete Antwort auf diese Ambivalenz liegt nicht in Ablehnung oder unkritischer Begeisterung, sondern in Kompetenz. Wer die Funktionsweise der Systeme versteht, ihre Ergebnisse überprüft, den Datenschutz beachtet und die ethischen Konflikte kennt, kann KI sicher, kritisch und nützlich einsetzen. Genau darin besteht der Beitrag, den Bildung und Prävention zu leisten haben: Menschen zu befähigen, diese Werkzeuge selbstbestimmt und verantwortungsvoll zu nutzen.

HINWEIS ZUR AKTUALITÄT

Der 31. Deutsche Präventionstag, auf den dieser Sammelband zurückgeht, fand am 13. und 14. April 2026 in Hannover unter dem Schwerpunkt „KI in der Prävention“ statt. Der vorliegende Beitrag wurde vom Autor Anfang August 2026 verfasst; die Inhalte wurden dabei gegenüber dem Kongressbeitrag noch einmal aktualisiert und auf diesen Stand gebracht. Das Feld der KI-Sprachmodelle entwickelt sich außerordentlich schnell: Die genannten Zahlen, Werkzeuge und Beispiele – insbesondere die Angaben in den Tabellen – geben den Stand von Anfang August 2026 wieder und werden voraussichtlich rasch an Aktualität verlieren. Wer aktuelle Zahlen und Inhalte wünscht, kann den Autor jederzeit kontaktieren; entsprechende Aktualisierungen werden gerne zur Verfügung gestellt. Die Kontaktdaten finden sich in den folgenden Autorengaben.

ÜBER DEN AUTOR

Mark Bottke (M.A. Erwachsenenbildung/Andragoge, B.A. Management im Sozial- und Gesundheitswesen), geboren 1971 in Berlin-Steglitz, ist Dozent, Trainer und Berater sowie Manager im Gesundheitswesen und als Lehrkraft am Gesundheitsverbund Landkreis Konstanz tätig. Unter der Marke MB-Trainings arbeitet er seit 1999 an der Schnittstelle von Kommunikation und Deeskalation, Beratung und Supervision sowie KI und digitaler Kompetenz und hat in dieser Zeit mit über 275 Organisationen und mehr als 7.500 Teilnehmenden gearbeitet. Als Geschäftsführer des Vereins Sicheres Freiburg engagiert er sich in der kommunalen Kriminalprävention. Als Kampfkünstler mit dem 1. Dan (Schwarzgurt) in Kung Fu und als Krav Maga Full Instructor verbindet er körperbezogene Trainingsarbeit – etwa zu Körpersprache und Status-Spielen – mit digitaler Kompetenz. Mit KI-Vorträgen ist er in sozialen Einrichtungen, in Dienstleistungsberufen sowie für namhafte Großunternehmen, etwa aus Telekommunikation und Luftfahrt, unterwegs. Ein Schwerpunkt seiner aktuellen Arbeit liegt auf dem sicheren, kritischen und praxisnahen Einsatz von KI in Bildungs-, Gesundheits- und Verwaltungskontexten. Kontakt: mark.bottke@mb-trainings.de, www.mb-trainings.de.

Literaturverzeichnis

- Anthropic. (2026). Statement on the US government directive to suspend access to Fable 5 and Mythos 5. <https://www.anthropic.com/news/fable-mythos-access>
- Deutsches Schulportal. (2025). AIS.chat, fobizz und Co: Diese KI-Tools können Schulen nutzen. <https://deutsches-schulportal.de/unterricht/fobizz-schulki-und-co-welche-ki-tools-koennen-schulen-nutzen/>
- Garland, A. (Regie). (2015). Ex Machina [Film]. Universal Pictures / A24.
- McCarthy, J., Minsky, M. L., Rochester, N. & Shannon, C. E. (1955). A proposal for the Dartmouth Summer Research Project on Artificial Intelligence.
- NewsGuard. (2025). Audit: Chatbots verdoppeln in einem Jahr den Anteil an Antworten mit falschen Informationen. <https://www.newsguardtech.com/de/ai-monitor/audit-chatbots-verdoppeln-in-einem-jahr-den-anteil-an-antworten-mit-falschen-informationen/>

- Otter.ai. (2025). Otter.ai caps transformational 2025 with \$100M ARR milestone, industry-first AI meeting agents. <https://otter.ai/press>
- Oura Health. (2026). Oura Ring 5. <https://ouraring.com/de>
- OWASP. (2025). OWASP Top 10 for LLM Applications – LLM01: Prompt Injection. <https://owasp.org/www-project-top-10-for-large-language-model-applications/>
- Presseportal. (2025). telli an weiteren Schulen im Einsatz: Schleswig-Holstein, Saarland, Bayern und Nordrhein-Westfalen führen KI-Chatbot ein. <https://www.presseportal.de/pm/180963/6172965>
- Scott, R. (Regie). (1982). Blade Runner [Film]. Warner Bros.
- Similarweb. (2026). Top generative AI chatbots & LLMs by market share. <https://www.similarweb.com/>
- Statcounter. (2025). AI Chatbot Market Share Worldwide. <https://gs.statcounter.com/>
- Stiftung Warentest. (2026). KI-Chatbots im Test: Perplexity schlägt ChatGPT und Meta AI. test.de. <https://www.test.de/KI-Chatbots-im-Test-Perplexity-schlaegt-ChatGPT-und-Meta-AI-6275046-0/>
- Straiker. (2025). Why 94% of AI agents are vulnerable to prompt injection – and what to do about it. <https://www.straiker.ai/blog/why-94-of-ai-agents-are-vulnerable-to-prompt-injection>
- TechCrunch. (2025). ChatGPT users send 2.5 billion prompts a day. <https://techcrunch.com/2025/07/21/chatgpt-users-send-2-5-billion-prompts-a-day/>
- TechCrunch. (2026). Zoom launches an AI-powered office suite, says AI avatars for meetings are coming soon. <https://techcrunch.com/2026/03/10/zoom-launches-an-ai-powered-office-suite-says-ai-avatars-for-meetings-are-coming-soon/>
- Timekettle. (2026). Timekettle kündigt für 2026 Durchbrüche im Bereich der KI-Dolmetschung an. PR Newswire. <https://www.prnewswire.com/news-releases/timekettle-kundigt-fur-2026-durchbruche-im-bereich-der-ki-dolmetschung-302652783.html>
- Turing, A. M. (1950). Computing machinery and intelligence. Mind, 59(236), 433–460.
- Zoom. (2025). Zoomtopia 2025: Zoom rolls out avatars and real-time translation for Workplace users. UC Today. <https://www.uctoday.com/unified-communications/zoomtopia-2025-zoom-rolls-out-avatars-and-real-time-translation-for-workplace-users/>